UNIVERSITÀ DI PARMA

UNIVERSITÀ DEGLI STUDI DI PARMA

Dottorato di Ricerca in Tecnologie dell'Informazione

XXXI Ciclo

Sentiment Analysis: from pre-processing to Applications in Online Communities

Coordinatore:

Chiar.mo Prof. Marco Locatelli

Tutor:

Chiar.mo Prof. Monica Mordonini

Dottorando: *Paolo Fornacciari*

Anni 2015/2018

Dedicated To
My Family
Letizia
Elia

Abstract

The high diffusion of social media is one of the most exciting novelty in these last years. Social media are not only used as a tool for messaging and sharing private things, but they are also used by people who want to share their opinion about some products or services. The huge amount of textual data produced by web social media has grown accordingly and there are obvious benefits for companies and governments in understanding what people think about their products and services, but it is also in the interests of public institutions to be able to collect, retrieve and preserve all the information related to specific events and their development over time.

Sentiment Analysis, which is the set of Natural Language Techniques for the identification and the categorization of opinions expressed in a piece of text, is of particular interest in order to determine attitudes towards a particular topic and can be successfully applied to the messages left in online social media.

However, most of the works regarding polarity classification usually consider text to infer sentiment and do not take into account that social networks are actually networked environments. For this reason, the combination of content and relationships is a core task of the recent literature on Sentiment Analysis.

Starting from the classical state-of-the-art methodologies where only text is used to infer the emotions expressed in social networks messages, this thesis presents two main contributions. The first contribution has been mainly focused towards some preliminary considerations for any kind of sentiment analysis: the accurate preprocessing phase of some available datasets, action never performed in a complete and accurate way in the relevant literature, the study and implementation of a novel and suitable polishing method based on an iterative learning approach and the comparison of different types of classifiers.

The second main contribution regarded the application of sentiment analysis to social networks in order to obtain a sort of combined approach: the network topology can contextualize the results of the Sentiment Analysis, while the polarity and the emotions expressed in the network can highlight the role of semantic connections in the hierarchy of the communities in the network itself. First, a sentiment has been associated to the nodes of Twitter graphs, showing the social connections, in order to highlight the potential correlations, i.e., similar ways to participate into a community. Then, sentiment analysis was applied to particular communities of Facebook, applying both automatic emotion detection and social network analysis techniques. This permitted to study how emotions are influenced by different kinds of relationships. Finally, after an up-to-date analysis of the state of the art for the problem of troll

detection, a systematic collection and grouping of features and a comparison among the different detected features with a machine learning approach, sentiment analysis was employed to detect malicious and anti-social behaviors in social networks, with the implementation of TrollPacifier, a novel holistic system for troll detection that demonstrated to reach a very high accuracy (95.5%).
The obtained results demonstrate that sentiment analysis can corroborate social network analysis and that together they can result a powerful tool to deepen the knowledge of online social network themselves.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Monica Mordonini for the continuous support of my Ph.D. study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of my research. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Michele Tomaiuolo, Prof. Agostino Poggi, and Prof. Stefano Cagnoni, for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

My sincere thanks also go to Laura, Giulio, Gianfranco, Alberto and Riccardo. Without their precious support, it would not have been possible to complete my research.

Finally, I would like to thank my family, Letizia, and Elia. They supported me spiritually throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sentiment Analysis (SA) is a discipline that has evolved, in the last few years, as a more and more important branch of text analysis.
Indeed, *sentiment analysis* is one of the most emerging research trends in Computer Science, a branch of opinion mining and computational linguistics, which regards computational methods to mine, evaluate, listen to, understand and process opinions, feelings, and reviews expressed in blogs, micro-blogs, forums, social networks and so on. Its primary purpose is to identify emotional states, to be an independent analysis tool and to enrich and complement other mining techniques, e.g., social sensing.

The correct and successful employment of sentiment analysis gives birth to several different issues, not so trivial to handle. For example, the general sentiment assessment may be divided into various steps. A first step is to differentiate objective from subjective considerations, then a second phase, also known as polarity detection, may regard to recognize positive, negative and neutral opinions about a particular topic [100]. A further subdivision may affect both positive and negative feelings; this is, indeed, what is appropriately called “sentiment analysis" as the previous steps mainly concern what is known as “opinion mining". As a matter of fact, in recent research works, SA goes beyond the concept of polarity, trying to identify the emotional status of a sentence, according to various classifications of effective knowledge [83, 118, 44, 23, 105]. Positive sentiments may have the following mood nuances: joy, friendship, hope, re-

lax, serenity, affection, sympathy and so on, while the negative ones may encompass anger, disgust, sadness, anxiety, depression, fatigue, fear, loneliness and the like. The granular detail, with which a sentiment model is described, is an essential factor affecting both complexity, derived from the human neural system, and accuracy of the obtained results, to correctly reflect the multifaceted sentiments of the real world.
Besides this, further issues in this analysis method concern the differentiation among the possible languages in which a word or a concept is written, as these may convey different feelings according to the various pronunciation of the vowels and the consonants. Some products/items may have different names across the world and so raise different emotions according to the word used and the culture of the provenance of the user.
The assessment of the sentimental influence of a post, a comment or an opinion, measured according to either the number of likes or re-tweets or in another way, is another task that should be carried out in more detail in this type of analysis.
The study of how the induced sentiments, produced by an influencer towards the influencees, comply or not with the original message is an important task to be carried on in many marketing campaigns, as well as in other fields, such as in the early discovering of abuses on the Internet, and it should accompany the assessment of various types of users: extroversive or introversive, agreeable or antagonistic, conscientious and unreliable, emotional or neurotic, open or closed to new experiences.
Finally, accurate sentiment analysis has also to deal with inter-sentiments inner relations and with some psychological constraints of the human mind, such as the law of sentimental inertia, the rule of origin asymptotically stability, the law of sentimental conflict, the principle of sentiment diffusion and the like.

Sentiment analysis is starting to play a crucial role in other fields as well, such as in managing customer relationships, detecting clients attitudes, positioning brands and products and developing effective marketing strategies. This is especially true when one considers that passions and emotions may drive customers' needs and willingness to buy items, even though often their cost is high and their necessity is disputable.
Nowadays, the application of SA ranges over several domains, from movie reviews to social networks, which are also proliferating in both usage and architectures [49, 47].

The demand for new techniques of SA is continuously growing, due to its inherent capacity of automatic evaluation, from both the academic and industrial points of view.

Social Media refers to web-based means of interaction among people, through which they create, share and exchange various types of information ranging from pictures to texts, from music to videos and so on, in a sort of virtual community. Some of the most popular are Facebook, Twitter, LinkedIn, YouTube, Instagram, Google+, Tumblr, Flickr, forums, blogs and so on.

Online Social Networks are a subset of Social Media and provide a collection of web-based services, to build a profile, interconnections of actors (individuals, groups, organizations, etc.) that usually share common and existent bounds in real life, such as friends, relatives, colleagues and the like [69]. The purpose of a social network may be multifaceted: interaction with friends and families, the creation of new business contacts, sharing photos and experiences, sharing emotions and feelings at any time and across distance, meeting new people, and so forth. Anyway, they are one of the most used ways people exploit to communicate with each other over physical boundaries and distances and can be roughly classified as user-oriented, where networking is the core application, and content-oriented, in which the sharing of images, opinions, videos and so forth, is predominant [15]. In the first group, the relationships, with known or potentially interesting people, are the real value and motivation that fosters the participation of the users [75], whereas in the second subset the interest is in what is shared and employable by the users themselves. Users of social networks may be seen as various subgroups of the whole of Internet users, or, better, identities, which can overlap throughout different social networks. We have underlined the concept of identities, rather than users, because multiple identities controlled by a single user may be expressed, e.g., deploying different email addresses.

Researches [121] showed that the usage of online social networks may help users to improve the quality of life by creating new relationships with new people who have similar interest, by tightening the existing relationships, and by expressing their sentiments, opinions, likes and dislikes. Such sentiments can increase a person's satisfaction in terms of self-esteem and belongingness and hence foster positive impacts

in psychological and social health [121].
Social networks are usually assessed in three ways [68]: structured-oriented, with emphasis on the topological connections, actor-oriented, with focus on the users' behaviors and actor-structure crossing, where the attention is on the interaction between connections and the users' actions.

It is evident that the "like never before" availability, provided by social networks, of information about single behaviors, common expressions, personal relationships, and so on, provides hints regarding a great range of phenomena and it affects various kinds of processes, be they political, social, commercial, educational, and so on.

Furthermore, the ever-increasing number of social media and social networks users provides more and more value and reliability to the statistical samples considered. For these reasons, the process of mining information about how people use their own time on the Internet and which kind of traces they leave, results to be very valuable and precious.

The growth of online social networks extends and improves the benefits, both for an individual and organizations, coming from the interactions among the users (the so-called social capital [48]) and much work has been done to try to model complex systems like social networks efficiently [111, 5]. The ability to retrieve and analyze large amounts of data, in particular, the chance to predict the collective decision by automatic data classification [18], has attracted the interest of marketing and politics.

The automatic classification of human activities is a well-known problem in different research areas [122, 87]. In the case of social-network analysis, Sentiment Analysis (SA) techniques [82, 94], as well as the study of the dissemination of information [46, 91], have been applied to the users belonging to a given network [4, 62].

Emotional states like joy, fear, anger, and surprise are encountered in everyday life and social media are more and more frequently used to express one's feelings. Thus, one of the main and most frequently tackled challenges is the study of the mood of a network of users and of its components (see, for example, [91, 62, 4]). In particular, emotion detection in social media is becoming increasingly important in business and social life [71, 118, 117, 105, 48].

In this thesis, we show some interesting results about sentiment analysis and the combination of sentiment analysis and social network analysis itself. As a matter of fact, the latter can contextualize the results of the former, e.g., to perform text classification and spam detection; while the polarity and the emotions expressed in the network can highlight the role of semantic connections in the hierarchy of the communities in the network itself.

As regards sentiment analysis we demonstrated in this thesis:

- the importance of a proper pre-processing phase, with particular focus on a basic cleaner, stemming and stop words removal, and the possibility to neglect the usage of a dictionary;
- the importance of polishing the dataset through a distant supervision approach and pruning techniques;
- the importance to use a hierarchical classifier rather than a flat classifier.

As regards sentiment analysis combined with social network analysis we showed in this thesis:

- a correlation between emotions expressed in the posts of a social network and the number of friends in the social network itself;
- the possibility to combine sentiment analysis and social network analysis to identify bad behaving users in social networks;
- the successful deployment of the combination of sentiment analysis and social network analysis to evaluate the reactions of people to political outcomes.

The structure of the rest of the thesis is the following: chapter 2, where the relevant literature about sentiment analysis and its application to social networks is reviewed; then two main parts describing the obtained results about sentiment analysis itself and about the application of sentiment analysis to social networks. Finally the conclusions summarize the whole work, pointing out also possible future developments.

Chapter 2

Related Works

2.1 About Sentiment Analysis

Due to the heterogeneity of usages of sentiment analysis itself, many different techniques were analyzed and implemented, to get increasingly accurate systems for a particular problem statement. Most of such techniques involve the use of *Machine Learning* (ML) classification algorithms, in particular, *Supervised Learning Algorithms*, i.e., methods that are used to train a classifier, whose aim is the association of an input with its related class, chosen from a certain set of classes. The training is done by providing the classifier with several examples of inputs and their related classes. Then, the system extracts a set of *features* (or *attributes*) from each of them, to become capable of recognizing the class of generic data, which can be of different types [87].
The performance of a classifier could be evaluated by different metrics, such as the *accuracy*, which is a measure of the correctness of a classifier, the *precision*, a measure of the ratio of correctly predicted positive observations to the total predicted positive observations, the *recall* or sensitivity, which measures the ratio of correctly predicted positive observations to the all observations in actual class, the confusion matrix, which is useful for the identification of the errors in the model classification and the like.

Indeed, Machine Learning algorithms need to work on data, appropriately processed by a set of operations which make assumptions and choices on the inclusion of features in text representations. This phase is a fundamental step for the whole system to obtain good results. Normally it includes methods for data cleaning and feature extraction and selection. A good overview of the steps and the most known algorithms for each step is explained in [77].

Thus, given a corpus of raw datasets, the first step of SA is the pre-processing of those data. Pre-processing involves a series of techniques which should improve the next phases of elaboration, to achieve better performances.

As illustrated in [60], online texts usually contain lots of noise and uninformative parts, such as HTML tags. This raises the dimensionality of the dataset and makes the classification process more difficult. The algorithms that are most used to polish and prepare data that come, for example, from Twitter messages, include the removal of punctuation and symbols, tokenization, stemming, and identification of stopwords as shown, for example, in [43] and [110].

Some of these techniques are exposed in the work of A. Balahur [11, 13], which concerns the problem of classification of Twitter posts, i.e., short sentences which refer to one topic. The author utilizes a series of pre-processing modules (such as emoticon replacement, tokenization, punctuation marks, word normalization, etc.) and describes these methods in detail. However, such methods are collected together before data classification, and the emphasis of the work is not on why or how each of these modules helps in improving the accuracy of the classifier. The work focuses on the classification of many types of sentiments, from positive, negative and neutral, to anger, disgust, fear, joy, sadness, and surprise, rather than on the effectiveness of the presented pre-processing techniques.

The work of A. Agarwal et al. [2], also based on Twitter data sets, proposes the use of emoticons as features and uses a dictionary of 8000 words associated with a pleasantness score from 1 (negative) to 3 (positive). Emoticons are divided into five categories (extremely-positive, positive, neutral, negative and extremely negative), and they gain a score, like other words. Then, all scores are added up and divided by 3. If the result is less than 0.5, then the sentence is classified as negative. If, on the contrary, it is

greater than 0.8, then the sentence belongs to the positive class. In all other cases, a neutral class is used. Basic cleaner, slang conversion and negation replacement are also used.
In the context of SemEval (Task 4)[1], for SA in Twitter, N. F. Silva et al. [116] analyze how much the accuracy of classification changes, using various algorithms: Naive-Bayes Multinomial (NBM), Support Vector Machine (SVM), AdaBoost with SVM, and AdaBoost with NBM. The authors showed that the use of AdaBoost provides good performance in the sentiment analysis (message-level subtask). In particular, in the cross-validation process, Multinomial Naive Bayes (MNB) has shown better results than Support Vector Machines (SVM) as a component for AdaBoost.

Considering the need for a large and clean dataset, distant supervision method has been shown to be an effective way to overcome the need for a big set of manually labeled data to produce accurate classifiers [56, 106]. Distant supervision is a semi-supervised method to retrieve noisy data, which are used to train traditional supervised systems. In [65] these methods are used to remove noisy data from automatically generated datasets of text (mentions) with good results. In particular, the results of this work show that a combination of mention frequency cut-off, Pointwise Mutual Information and removal of mentions which are far from the feature centroids of relation labels is able to significantly improve the results of two relation extraction models. A survey of dataset pruning methods for distant supervision in sentiment analysis is exposed in [109]. In this paper, authors have categorized the approaches into three categories: First, models that are based on the principle that it is necessary and sufficient that at least one context expresses a fact in the knowledge base. Second, hierarchical topic models that estimate different distributions for background, relation-specific, and pair-specific contexts. Third, an approach that employs argument correlations between patterns.

Once that the available dataset has been pre-processed and polished, SA needs proper classifiers and mining techniques to perform its tasks in an effective way.
Recent and comprehensive surveys of sentiment analysis and the main related data analysis techniques can be found in [82, 94]. Authors describe automatic systems

[1] http://alt.qcri.org/semeval2016/task4/

and datasets commonly used in sentiment analysis, summarize several manual and automatic approaches to creating valence and emotion association lexicons.

As concerns the used tools, hierarchical classifiers are widely applied to large and heterogeneous data collections [42, 115, 1]. Essentially, the use of a hierarchy tries to decompose a classification problem into sub-problems, each of which is smaller than the original one, to obtain efficient learning and representation [76, 8].
Moreover, a hierarchical approach has the advantage of being modular and customizable, with respect to single multi-class classifiers, without any loss of representation power: Mitchell [92] has proved that the same feature sets can be used to represent data in both approaches.

Emotions in tweets are detected according to a different approach in [3]. In this work, polarity and emotion are concurrently detected (using, respectively, SentiWordNet [44] and NRC Hashtag Emotion Lexicon [95]). The result is expressed as a combination of the two partial scores and improves the whole accuracy from 37.3% and 39.2% obtained, respectively by independent sentiment analysis and emotion analysis, to 52.6% for the combined approach. However, this approach does not embed the *a-priori* knowledge on the problem as effectively as a hierarchical approach, while limiting the chances to build a modular, customizable system.

Some tools providing more specific classifications than the simple positive or negative polarity of the classical SA have also been developed in practice [118, 117, 105, 71]. In [3] emotion analysis on brand tweets are conducted using both approaches of SentiWordNet [10] and NRC Hashtag Emotion Lexicon [95], without relying on any a-priori knowledge. In [119], Plutchick's wheel of emotion [104] is used to treat the inherently multi-class problem of emotion detection as a binary problem, for four opposing emotion pairs.

2.2 About the application of Sentiment Analysis to Social Networks

In this section we report the related works and the state of the art about the combination of different techniques in order to make a deeper and multifaceted Social Networks

Analysis (SNA), in terms of emotions expressed and their relationships with social media elements (i.e, publishing time and date) and with social aspects concerning members activities in a social network group (i.e., interactions and friendships).

SNA has the objective to model social structures with different properties, starting from the mathematical theory of graphs and the use of matrix algebra, and is often augmented though computer-based simulations [49]. In order to make social networks more intelligent and flexible, a deeper analysis of effective knowledge could be incorporated [49]. In some case an ontology-driven approach is used [12, 118, 10]. Some recent studies about American candidates are important for understanding how public sentiment is shaped in social networks and its polarization [94]. The contribution in [4] exploits geospatial information related to tweets for estimating happiness in Italian cities. However, the techniques used generally in SA and Text Classification must be adapted to the maximum number of characters that some social networks feature for their Status Update Messages (SUMs), and this opens the way for new issues [2, 78, 73, 128].

Moreover, the advent of social media has radically changed the way in which chronic patients looking for information and support share their condition. In fact, online social networks provide more and more medical data because patients often share personal clinical information, with the aim of receiving emotional and practical support. In recent years, the interest in patients' opinions and feelings expressed in web communities has considerably increased. One of the biggest challenges is to get a clear understanding of the patients' condition. In [59] the authors identify the 15 largest Facebook groups focused on diabetes management, analyzing 690 comments from wall posts written by 480 unique users. The work aims at identifying, with a traditional manual method of content analysis, the main topics of the discussions. Since this approach is not scalable when trying to analyze huge quantities of data, it is necessary to introduce automatic analysis, based on machine learning algorithms. In [58] authors analyze different sources of patients' information (social media, blog, patients networks) to detect poor quality healthcare, using sentiment analysis and natural language processing. Sentiment Analysis using machine learning algorithms represents an automatic way to analyze sentiments, emotions, and opinions from

written language [82, 94, 71] and it is becoming increasingly important in business and social contexts [48]. Another interesting case is [108], where the authors, using Sentiment Analysis techniques, try to understand what the medical community could learn from the information that is shared on the Chron disease web community. The authors analyze patient's opinions about therapies and drugs, studying the most debated topics.
Interests have also grown towards the analysis of the social mechanism and dynamics inside these patients communities. For example, in [86] the authors compare Facebook pages on different chronic illnesses and also looking at how patients and other stakeholders talk about the same chronic diseases on Twitter, while in [112] an evolutionary metaheuristics approach has been used to identify communities or subsets of users in a Facebook group of patients.

The Computer-Mediated Communication (CMC) provided by social networks can also furnish varying degrees of anonymity that can encourage a sense of impunity and "freedom" from responsibility for users. This whole scenario has led to the development of a widespread phenomenon that occurs within the CMCs, known as trolling.
The first references to the use of the word troll on the World Wide Web have been found in Usenet, a forum community popular in the eighties. A troll is generally defined as an individual who is marked by a negative online behavior [61, 32], or as a user who initially pretends to be a legitimate participant, but later attempts to disrupt the community, not necessarily in a blatant way, but with the effect of attracting the maximum number of responses [40]. Trolls are also described as individuals who derive pleasure from annoying others [74], and, in fact, recent researches have discovered that sadism is closely associated with those who have trolling tendencies [19].

The contribution in [17] highlights the connection between dark personality traits and engagement in harmful online behaviors.

A troll seeks to cheat a person or a whole community [96]. In sociology, the term has become synonymous for all negative online behaviors, but it is necessary to recognize each one by giving them a definition to understand and face the online

trolling phenomenon in a systematic way.

Studying the behavior of some users within the virtual communities of Usenet, Hardaker [61] has found that the act of trolling is manifested through four interrelated ways:

- *Deception*: a troll will try to disrupt the group, trying to stay undercover; for example, when a troll intentionally disseminates false advice [40].
- *Aggression*: a troll that is searching for a conflict, can use a provocative tone towards other users.
- *Disrupt*: it is the act of causing a degradation of the conversation without necessarily attacking a specific individual.
- *Success*: often a troll is acclaimed by users for his degree of success, so trolling, despite being a nuisance for users, may end up at the center of attention of the group.

It is clear that trolling is a more complex problem than just the source of provocative attacks. Although the concept may seem tied to the meaning of some words like rudeness, arrogance, impertinence, and vulgarity, they do not provide an accurate description, since typically trolling consists in keeping hidden the real intent of causing problems. The contribution in [20] shows the characteristics of troublemakers in online social. This study provides significant multilevel support for the association between socio-demographic factors, communication patterns and structural network characteristics on one side, and troublesome contacts in online networks on the other.

A recent study states that anyone can become a troll: in fact, their predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling [26].

These practices are often tolerated, in line with a prevailing attitude on the Internet that considers offensive speech as a manifestation of freedom of expression [103]. In less vulnerable communities, with more experienced or emotionally detached users, some episodes can also be seen as playful actions. However, inexperienced

or vulnerable users of online communities may feel trolling particularly painful, distressing and inexplicable.

To counter these actions, some services implement identity verification processes [63]. Nevertheless, the propensity to trolling seems to have become more widespread recently [22]. A case study analysis of the behaviors and strategies of a group of alleged Twitter trolls is presented in [120]. In extreme cases, anti-social online behavior has also led to suicides of adolescents [51]. Thus, it is not surprising that this growing phenomenon is alarming the social network operators [57].

Even when trolling does not come as a direct attack, it can be a threat because it can manifest itself with subtler ways, for example as a means to try to manipulate others' opinions. The rise of the Internet has allowed corporations and governments to disseminate false rumors freely, or to use other dishonest practices to polarize opinions [35]: it has been shown that a user's opinion can be influenced by other users' comments [34, 72].

Considering its diverse motives and forms, trolling represents a vexing problem in CMC, because it hinders the ordinary course of a conversation. Indeed, user contributions in the form of posts, comments, and votes are essential to the success of an online community. However, with such a high degree of desired participation, excluding individuals with rude online behaviors, as trolls, can lead to a perception of excessive control and censorship, trigger side effects, impede effective community development. Thus, the goal to protect discussion threads from trolling has to be accurately balanced with a certain level of tolerance, for avoiding unnecessary interruptions and facilitating the integration of novice and uneducated users.

Usually, online social networks rely on moderators for banning malicious users. In many cases, also common users are provided with the option to flag inappropriate posts and mute users. However, this kind of manual solution has some major drawbacks, including a delay of actions, subjectivity of judgment and scalability [99]. Thus, it is necessary to augment the process through some automatic mechanisms. However, to create such a complex system, it is necessary to take into account the most distinguishing features of online trolls. Few research works consider some different aspects of online trolls. In a study on anti-social behavior in large online discussion

communities (CNN.com, Breit- bart.com and IGN.com), some general tendencies have been observed [27]. The analysis focuses on the users subsequently banned by the moderators, defined as "Future-Banned Users" (FBUs), and confronts them with more civil users, defined as "Never-Banned Users" (NBUs). Analyzing their behavior before being banned, FBUs show a tendency to write comments which are difficult to understand, often off-topic and with an adversarial language [37]. They tend to focus on few discussion threads, but they contribute with more posts per thread and they also receive more answers than average, suggesting that success in attracting attention can be synonymous with abnormal behaviors. Furthermore, FBUs have a high rate of post-cancellation (by moderators) and signaling (by other users), increasing over time. The described system can predict when an individual will be banned, with over 80% of accuracy, analyzing four sets of characteristics: post content, user activity, reactions of the community, moderator's actions. A similar study [89] has been conducted on the community of an online newspaper (Dnevnik.bg). Authors have derived specific metrics, including community rating, consistency with the topic, the order of comments, answers, time of the day. A definition of troll as somebody who was called such by other people was used in [90] to predict, in news community forums, whether a troll writes a comment or not. In this work, most of the features are based on textual attributes, but they are evaluated in a good methodical way. In fact, the authors report the results of classifiers trained (i) using all features, as well as (ii) excluding one individual feature group.

The majority of research works in this area focus their analysis on few homogeneous features. To study this large variety of proposed analyses and research works systematically, we have identified six main types of approaches. For each type of approach, we have defined a group of features, using ideas proposed in previous researches as well as new ones. This systematic survey of the scientific literature is not intended as a mere study or a reasoned comparison, but instead, it is intended as the first step for creating an online automatic troll detection system.

Some research studies apply sentiment analysis to the problem of troll detection. For example, in [127], sentiment analysis is applied on the Twitter social network, and it is used to identify political activists hostile to other parties and to evaluate

the degree of conflict between two different factions, during the electoral period in Pakistan. The researchers use a tool called SentiStrenght, which estimates the "force" of a sentiment (either positive or negative). In [84], another study is reported, likewise characterized by the analysis of political discussions on Twitter, which tries to spot the malevolent users through the content of their tweets. Using a similar approach, the VaderSentiment library [55] is based on a lexicon sensitive to both the polarity and the intensity of sentiments of words. It has been validated by multiple independent human judges and is tailored especially for microblog-like contexts. Nevertheless, according to its authors, it is also applicable in other domains. Another proposed paradigm for text analysis in this field is "sentic computing" [22]. This paradigm is more focused on semantics rather than syntax, and it is more inclined to evaluate the sense of the text, including what is expressed implicitly. This model is not shaped on static learning models, but it uses tools based on domain-specific ontologies.

In [6], an emotion detection system is described. The system is based on a hierarchy of classifiers, at three levels. The classifiers at the three levels distinguish, in order: objective / subjective tweets; positive/negative tweets (among the subjective ones); tweets expressing fear/ anger/sadness (among negative tweets), or love/ joy/surprise (among positive tweets).

The work illustrated in [39] tries to estimate, solely with metadata, the presence of trolls inside the reddit.com portal, and highlights some characteristics according to the criteria set out above. All the obtained information is collected in attributes of instance variables used to train a Support Vector Machine classifier. Once tested, it has shown a good accuracy of about 70%. The results show that the approach based exclusively on metadata is less accurate than the ones based on the sentiment analysis, but a combination of the two could bring benefits to both methods, like, for example, it happens in [113].

The frequency of publication has been related to the quality of online discussions by various studies. In [27], the features of users later banned from some large websites are studied. In addition to the kind of produced text, also patterns of activities are observed. It is found that useful features, to distinguish future banned users, including the frequency of some activities, as the number of posts and comments per day.

In [36], newsroom interviews, reader surveys, and moderators' choices are used to characterize the comments published on a newspaper website. It is found that the frequency of commenting is a valuable indicator of low-quality discourse.

In [90], authors describe two classifiers: one for detecting "paid trolls", who try to manipulate a user's opinion, and one for detecting classical "mentioned trolls", who offend users and provoke anger. Among many features regarding sentiment and text analysis, based on lexicons and bag of words models, they also consider some metadata, including the publication time. In particular, they distinguish a worktime period (9:00-19:00h) and a nighttime period (21:00-6:00h). They also distinguish workdays (Monday-Friday) and weekend days (Saturday and Sunday). This kind of feature is found to have the most significant impact on accuracy, according to this study.

Various approaches have been studied to carry out troll detection through the evaluation of the textual content of online messages. Some studies are based on the evaluation of the ARI (Automated Readability Index) of published texts since it has been shown that a troll is more likely to write in a less comprehensible language compared to a normal user [27]. According to [39], a troll is more likely to write short comments, maybe because he writes faster replies compared to a non-malevolent user that writes more elaborated and longer sentences.

Other studies attempt to bring the troll identification problem to a higher level of analysis, studying not only individual messages but the entire discussion about the topics. This hybrid approach incorporates some of the techniques described in the previous subsection, but also adds new information obtained from the context in which the messages are integrated. Among them, [33] adopts a combination of metrics of a statistical and syntactic nature, and other elements related to the users' opinion: some of these measurements are similar to the ones already treated. Others manage to summarize more general properties of the discussion, like the number of references to other comments, how many times a determinate post is mentioned in the topic and the degree of similarity between the terms involved in the thread, which is a measure also used in other studies and obtained thanks to the cosine similarity [27, 89]. The approach conceived by [37] is made by evaluating the problem from the same point

of view, but using different concepts. It is based on the Dempser-Shafer theory [53], a generalization of the Bayes' probability concept, that turns out to be a very useful tool when it comes to imprecise and uncertain information, like the ones provided by the users of these environments. The study underlines how it is possible to characterize messages according to their apparent rationality, their degree of controversy and their relevance for the topic of discussion.

The necessity for integration of user level metrics for the problem of troll detection has emerged from various research works.

In [27], authors focus their efforts on the extraction of users' general data. The aim is to study the most significant parameters for the characterization of a troll, thus obtaining a better perspective of troll behavior. In [107], various metrics are described, to measure a user's involvement in the platform and the nature of his/her participation. Some of the described metrics aim at distinguishing active users from passive ones, by comparing the number of original tweets and replies produced, with the number of retweets, quotes and likes.

In [31, 29], the different problems emerging from the interactions of users with online bots are tackled. In [31], an approach inspired by the biological DNA is applied to the analysis of users' behavior on social networks. In this case, sequences are constituted by codes representing different types of social actions, namely comments, likes, shares, and mentions. While the particular behaviors of bots and trolls may largely differ, both aim at diverting attention from the discussion topic. Thus, detection methods developed for one kind of abusive behavior may also prove useful for the other one.

The community-level approach tries to solve the problem of troll detection through the study of the relationships within the online community, using the methodologies of social network analysis. To our knowledge, the first study which explores this field is reported in [80]. In the study, troll detection is just a part of a comprehensive analysis of Slashdot Zoo, a portal that allows each user to label others as friends or foes. Thanks to this peculiarity, the social graph has some links with negative weights, which represent distrust and are useful to identify unpopular users.

For troll detection, the most useful metrics are obtained through a variation of the

Page Rank algorithm, taking into account negative weights, and by the raw number of foes of a node.

In [107], a modified version of the Hirsch Index is proposed for measuring the influence of a user. The Hirsch Index (h-index) is used in the research community to evaluate the scientific production of a scholar, by the received citations. In the context of Twitter, it can be defined as the largest number n, such that n tweets of a user have been retweeted or liked at least n times.

In [79], authors explain how to transform any social network in one with "friends and enemies". As a result, several solutions to troll detection based on this approach were born. For example, in [79] researchers try to improve this method by implementing an algorithm that, at each iteration, reduces the size of the social network by eliminating all edges that are unnecessary for the analysis and focusing more on the types of "attacks" adopted by trolls. Instead, the work shown in [99] evaluates how it is possible to use the propagation of trust and distrust for measuring the reliability of a node.

Especially in the case of propaganda agents and opinion-spreading trolls, links to external content, like images, videos, and articles, pay an important role [9]. This can be the case of paid trolls, political activists, influencers and advertisers. Advertisements of this kind include links to external content, but also to groups, pages, and hashtags, often used to identify and mount viral campaigns. In fact, in recent years, social media are increasingly being used for creating coordinated and multi-faceted campaigns [67, 30]. Those activities include the role of human influencers, both willing and paid, troll users who try to disrupt the discourse of adversaries or to attack opponents personally, bots, external content creators and news outlets. Thus, the presence of many forms of content advertisement can be taken into consideration for detecting and managing anti-social behaviors, in general.

Part I

Sentiment Analysis

Chapter 3

Results about Sentiment Analysis

Considered the importance of Sentiment Analysis, as highlighted previously in the thesis, this chapter focuses on the creation of the best conditions, or pre-conditions, to perform a successful SA process. In particular, different pre-processing and polishing algorithms to create a good training dataset for Sentiment Analysis have been compared, and we evaluated which type of classifier permits to achieve the best performances.

First of all, we performed a comparison between pre-processing techniques over data sets usually employed in Sentiment Analysis. The purpose of this comparison was to evaluate which techniques were the most effective, but also to find out the reasons why the accuracy of the Sentiment Analysis improves in the presence of particular pre-processing methods. This was achieved by means of a precise analysis of each considered method.

A second step regarded the devising and the implementation of an iterative learning approach, which combines distant supervision with dataset pruning techniques. In particular, we applied a classifier, trained on raw data obtained from different Twitter channels, to the same original data set for removing the most dubious instances automatically. This approach produced a more polished training set for emotion classification, considering Parrot's model of six basic emotions.

A final contribution concerned a comparison between two approaches to emotion

classification in tweets, taking into account six basic emotions. Training data sets have been first collected from the web and then automatically filtered to exclude ambiguous cases, using an iterative procedure. Then, two approaches have been compared. The first one is based on a direct application of a single *flat* seven-output classifier, the second one is based on a *three-level hierarchy* of four specialized classifiers, which reflect the a-priori relationships among the target emotions. The described results demonstrated that the a priori domain knowledge embedded into the hierarchical classier makes it significantly more accurate than the flat classifier.

3.1 Comparison between Pre-processing Techniques for Sentiment Analysis

This section provides information about the compared algorithms and the techniques, as well as the results and the findings of the comparison. In particular, the different modules, implemented in Python version 2.7, that have been used in this pre-processing research are described.

3.1.1 Pre-processing techniques

The pipeline of the comparison is organized in the following way. First, we obtained the 2015 and 2016 data sets (both training and test) of Twitter Sentiment Analysis from **SemEval**. The training sets are then subject to the various pre-processing techniques taken into consideration in this research and described in the following. After the text of each instance of a set has been pre-processed, the resulting sentences (the *cleaned tweets*) become the instances of a new training set. Then, such a data set is used to train a classifier, and the corresponding test set is classified through Weka. Finally, the accuracies of the classifiers obtained from different pre-processing modules are compared with each other, to evaluate the efficiency and effectiveness of each technique. The whole pipeline is depicted in Fig. 3.1.

The classifier is made by using Naive-Bayes Multinomial (NBM) method, i.e., a machine learning algorithm that gives rise to a probabilistic classifier, which works on

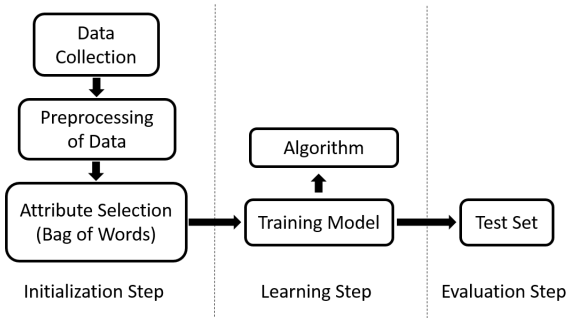


Figure 3.1: Steps to train a classifier for sentiment analysis.

the basis of the Bayes Theorem, with the strong assumption that features are mutually independent. Let X = $(x_1, \ldots, x_n)$ be the feature vector of an instance in the data set, that is, a binary vector that takes into account the presence of a feature in that instance, and let $C_1, \ldots, C_K$ be the possible outputs (classes). The problem is to gain the posterior probability of having the class C_k as output, given the feature vector X, and given the prior probability $p(C_k)$ for each class. Thanks to the Bayes Theorem and the independence between features, the probability that needs to be estimated is the conditional $p(X|C_k)$, and then a classifier is trained with a decision rule, such as the Maximum a Posteriori (MAP) rule. In summary, the probabilistic model of NBM can be expressed in terms of the following formula:

$$p(X|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i}$$

where $X = (x_1, \ldots, x_n)$ is the feature vector, p_i is the probability that the feature i appears, C_k is a class and p_{ki} is the probability that feature i occurs in the class C_k. Then, Information Gain (IG) is the algorithm used for feature selection. It evaluates the presence or absence of a feature in a document by measuring its probability of belonging to a class. The amount of information needed to exactly classify an instance D is defined recursively as follows:

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D_i|} \cdot Info(D_j)$$

where the instance D is divided by some feature attribute $A = \{a_1, \ldots, a_v\}$ into sub-instances $D_1, \ldots, D_v$.

3.1.1.1 Basic Operation and Cleaning module

This first module of the pre-processing manages basic cleaning operations, which consist in removing unnecessary or disturbing elements for the next phases of the analysis and in the normalization of some misspelled words. In order to provide only significant information, a clean tweet should not contain URLs, hash-tags (e.g., #happy) or mentions (e.g., @BarackObama). Furthermore, tabs and line breaks should be replaced with a blank and quotation marks with apexes. This is useful to obtain a correct elaboration by Weka (i.e., not closing a quotation mark causes a wrong reading by the data mining software causing a fatal error in the elaboration). After this step, all the punctuation is removed, except for apexes, because they are part of grammar constructs such as the genitive. The next operation is to remove the vowels repeated in sequence at least three times because by doing so the words are normalized: for example, the words *cooooool* and *cool* will become equals. Another substitution is executed on the laughs, which are normally sequences of "a" and "h". These are replaced with a "laugh" tag. The last step is to convert many types of emoticons into tags that express their sentiment (e.g., :) → smile happy). The list of emoticons is taken from Wikipedia[1]. Finally, all the text is converted to lower case, and extra blank spaces are removed.

Finally, all the text is converted to lower case, and extra blank spaces are removed.

All the operations in this cleaning module are executed to try to make the text uniform. This is important because, during the classification process, features are chosen only when they exceed a particular frequency in the data set. Therefore, after the basic pre-processing operations, having different words written in the same way helps the classification.

[1] http://en.wikipedia.org/wiki/List_of_emoticons

3.1.1.2 Emoticon module

This module reduces the number of emoticons to only two categories: *smile_positive* and *smile_negative*, as shown in Table 3.1.

Table 3.1: List of substituted Emoticons

smile_positive	smile_negative
0:-)	>:(
:)	;(
:D	>:)
:*	D:<
:o	:(
:P	:|
;)	>:/

Table 3.2: Likelihood of some Emoticon

Features	$P(X\|C_{pos})$	$P(X\|C_{neg})$
:)	0.005107	0.000296
:(	0.000084	0.001653
:*	0.001055	0.000084
;)	0.000970	0.000084

This is done to increase the weight of these features in the classification phase and to reduce the complexity of the model. In Table 3.2 and Table 3.3 it is possible to notice how much the likelihood of the features change.

Table 3.3: Likelihood of Smile_Positive and Smile_Negative

Features	$P(X\|C_{pos})$	$P(X\|C_{neg})$
smile_positive	0.007320	0.000718
smile_negative	0.000336	0.002283

3.1.1.3 Negation module

Dealing with negations (like "not good") is a critical step in Sentiment Analysis. A negation word can influence the tone of all the words around it and ignoring negations is one of the leading causes of misclassification.

In this phase, all negative constructs (*can't*, *don't*, *isn't*, *never*, etc.) are replaced with "not".

This technique allows the classifier model to be enriched with many negation bigram constructs that would otherwise be excluded due to their low frequency. Table 3.4 shows some examples of extracted features and their *likelihood*.

Table 3.4: Example of Features Extracted

Features	$p(X\|C_{pos})$	$p(X\|C_{neg})$
not wait	0.002345	0.000304
not miss	0.000651	0.000043
not like	0.000004	0.000391

3.1.1.4 Dictionary module

This module uses the external python library PyEnchant[2], which provides a set of functions for the detection and correction of misspelled words using a dictionary.

As an extension, this module allows one to replace slang with its formal meaning (e.g., l8 → late), using a list. It also allows one to replace insults with the tag "bad word".

The motivation for the use of these functions is the same as for the basic pre-processing operation, i.e., to reduce the noise in text and improve the overall classification performances.

3.1.1.5 Stemming module

Stemming techniques, employed in this module, put word variations like "great", "greatly", "greatest" and "greater" all into one bucket, effectively decreasing entropy and increasing the relevance of the concept of "great". In other words, stemming allows us to consider in the same way nouns, verbs, and adverbs that have the same radix.

This method is already implemented in Weka, and the algorithm in use is *IteratedLovinsStemmer*[3].
As in the case of emoticons, with the use of this technique, it is possible to combine features with the same meaning and reduce the entropy of the model.

3.1.1.6 Stopwords module

This module addresses words which are filtered out in the pre-processing step. These words are, for example, pronouns, articles, etc. It is essential to avoid having these words within the classifier model because they can lead to less accurate classification.

[2]http://pythonhosted.org/pyenchant

[3]weka.sourceforge.net/doc.dev/weka/core/stemmers/IteratedLovinsStemmer. html

3.1.2 Results

The data set is composed of the training and test sets.

Table 3.5: Data set table

Data set	**Positive**	Negative	Total
Training set	1339	1339	2678
Test set 2016	623	169	792
Test set 2015	343	174	517

The training sets are those provided by SemEval, with a little revision: neutral sentences are removed, to focus only on positive and negative ones. Furthermore, in the training set, there are more positive sentences than negative ones. Excess positive ones have been eliminated because they distort the Bayes model.

In the executed tests, the features collected have a minimum presence in the text that is greater than or equal to 5. The Ngrams used are only unigrams and bi-grams. Before starting the simulation with the test set, 10-fold cross-validation is carried out. In particular, the optimal length of N-grams to potentially consider as features was searched. In Figure 3.2, it can be observed that accuracy nearly peaks at N-gram = 2. Longer sequences increase the complexity of the training phase, without giving a significant improvement in the result.

Moreover, the total number of features to consider has been analyzed. This parameter does not provide a monotonic improvement to the classifier quality. Instead, it peaks out at around 1500 features.

At first, the executed simulations compare a *no pre-processed* file vs. *basic cleaned* file. As shown in Table 3.5, the resulting accuracy is strongly increased using the cleaned file. Given the importance of the basic cleaner, we decided to use it in every case, together with another pre-processing module, to evaluate their contribution together.

Stemming increases the performance because it groups words reduced to their root form. It allows many words to be selected as useful features for the classification phase. In fact, it modifies the weight of a feature, usually increasing it.

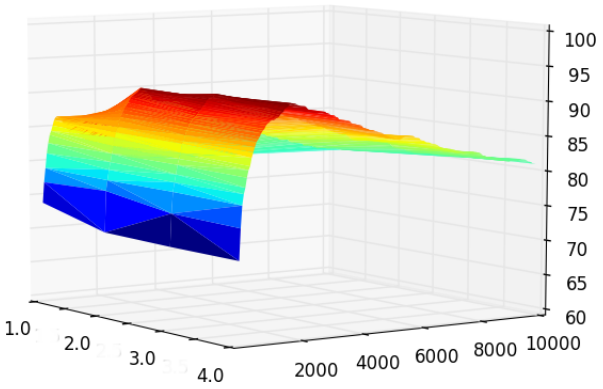


Figure 3.2: Optimization of system parameters.

Stop-word removal enhances the system because it removes words which are useless for the classification phase. As a common example, an article does not express a sentiment, but it is very present in the sentences.

Table 3.6: Result classification table

Technique	**# Max Features**	CV folds 10 [%]	Test 2016 [%]	Test 2015 [%]
No preprocess	1800	78,08	65,65	69,05
Basic	2000	80,05	65,40	74,08
Basic + Stemming	2200	**80,84**	**68,68**	**76,40**
Basic + Stopwords	1800	80,32	65,27	74,85
Basic + Negation	2000	80,40	65,65	75,04
Basic + Emoticon	2000	80,13	65,98	74,66
Basic + Dictionary	2000	78,00	64,39	75,82
All	2000	80,40	64,89	75,82
All Without Dictionary	2100	80,76	65,78	75,04

As a notable result, it is interesting that using a dictionary did not enhance the performance in the considered tests, but it increased the elaboration-time needed for cleaning raw data. There is also an improvement in the accuracy of the classifier between the two SemEval test-sets (2016 and 2015). However, this is only because

there are fewer sentences in the last test-set, with a correspondingly lower probability for the classifier to make mistakes.

3.2 Creation of a polished training set for sentiment analysis in an automatic way

This section describes the research carried out to create a proper training set for sentiment analysis of tweets from Twitter in an automated way. In fact, the performances of an automatic system for emotion analysis are mainly affected by the quality of the data set used to train it, but a few publicly available, reliable and manually annotated datasets are described in the scientific literature, and they are addressed only to valence (polarity) classification.

Given the high costs required for manually annotating a training set, we decided to devise and use an automated distant supervision approach. This approach was easily implemented because different users of Twitter tend to label their emotional states with specific hashtags corresponding to them.
The distant supervision approach has the advantage of allowing the collection of a solid training set, in a short time. However, its main disadvantage is the lack of control over the way people decide to label their tweets, resulting in noisy data. In the light of this situation, a distant supervision has been combined with an automatic data set pruning technique, that will be described in the following. In order to evaluate the effectiveness of the data set pruning phase we trained, in the same way, but with different training sets, a number of seven-outputs "flat" classifiers:

- **A raw classifier**: trained on the training set collected **using distant supervision without applying dataset pruning**;
- **A set of six improved classifiers**: trained on the training sets obtained from the **dataset pruning phase** executed with different thresholds;

and compared them on the same manually-annotated test set. We underline the importance of having a manually-annotated test set, in order to actually measure the validity of the studied approach.

Table 3.7: Hashtags selected for each sentiment.

Sentiment	Hashtags
Joy	#joy, #happiness, #happy, #joyful, #blessed, #smile, #goodvibes, #proud
Love	#love, #loveofmylife, #fiance
Surprise	#surprisesurprise, #wtf, #omg
Anger	#fuckyou, #pissedoff, #angry, #furious, #fuckoff, #annoyed, #stfu
Sadness	#sad, #sadness, #sosad, #disappointed
Fear	#terror, #scared

3.2.1 Training set creation

To implement the distant supervision approach, the Twitter REST API has been employed to download tweets containing some given hash-tags, corresponding to Parrot's primary sentiments, and other terms selected by an empirical study of tweets. In creating this dataset, we relied on the fact that Twitter's users when expressing emotions, add specific hash-tags corresponding to their emotions. To identify the most popular hashtags used to express a given sentiment, a set of tweets has been downloaded and manually searched for hash-tags used in a consistent way. We decided to download more hash-tags for each emotion to represent all possible different facets. The selected hash-tags for each emotion are presented in table 3.7.

Since the objective class is considered in the task of polarity classification and considering that there are publicly available datasets for this field, we have decided to collect the instances relative to the "objective" class from these sets. The datasets we have chosen are "SemEval-2013 Task #2" [97] and "Emotweet-28" [125].

The **raw training set** refers to the set of tweets downloaded using the hash-tags presented in Table 3.7, those collected for the objective class and the corresponding labels obtained as previously described. It has been essential to proceed with a pre-processing stage, similar to the one described in 3.1:

- tweets are cleared from elements with no emotional meaning, such as hashtags, user references, punctuation or retweet information;

- tweets are cleared from links;
- repetition of tweets are removed;
- emoticons and contractions are replaced with their textual extension;
- keys used to download the tweets are removed;

After these operations, the raw training set is composed of 42533 instances equally distributed within each class.

In the following, we will present the approach used to train the considered classifiers and the data set pruning technique we used. Starting from the raw training set, we will describe the algorithms and tools used to derive the raw classifier. Then, we will describe how this classifier has been used to derive the **filtered training sets**, starting from the raw training set.

3.2.2 Classification

The considered classifiers are the following: one from the raw training set and many others from the filtered training set, in a scheme known as "dogfood learning". In fact, following the "eat your own dogfood" principle, the classifier obtained from the raw training set has been then applied to the same initial raw data set, to filter out the more dubious instances automatically. As it will be described later, we have been able to filter out dubious instances at different levels, and hence obtain a different classifier for each of the "cleaning levels" considered. All classifiers have been trained with the same approach: using the *Naive Bayes Multinomial* algorithm (in particular, the implementation provided by Weka). To define the features of the training set, the *String to Word Vector* algorithm has been used, that turns a string into a set of attributes representing word occurrences. However, it is important to use not only uni-grams (single word) but to extend the representation to n-grams (set of maximum "n" words). To select the features that are more relevant for the training sets, we have used the *Information Gain* algorithm (also in this case through the implementation provided by Weka).

For each training set, a preliminary phase has been dedicated to optimize the parameters representing the number of features and of n-grams to be used. The phase started from a grid of pairs *(n-grams, number of features)* and used *cross-validation* to estimate the quality of classifiers configured with the parameters defined by these pairs. Then, we used the pair that returned the best results. Figure 3.3 shows the case of the **raw classifier**; it can be noted that the accuracy peak corresponds to n-grams = 2 and number of features = 6760.

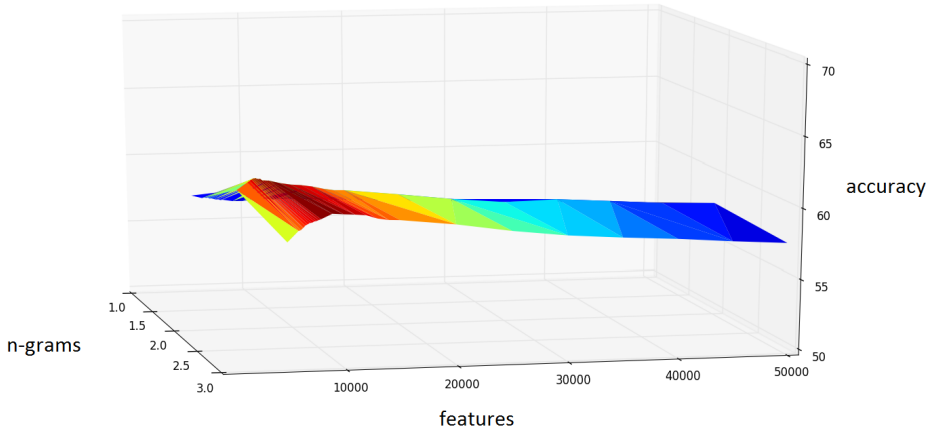


Figure 3.3: Parameters optimization.

3.2.3 Dataset pruning

The basic assumption underlying the considered dataset pruning scheme is that the most uncertain instances, contained in the raw training set, represent only a fraction of the ones that are correctly classified. This hypothesis has been considered true since the results of the work described in [93] show that the instances obtained by distant supervision have similar quality to annotations of trained human judges.

Figure 3.4 summarizes the data set pruning scheme used in this section, which is composed of the following sequential steps:

1. Training a classifier (which we call the "raw classifier") using the whole raw

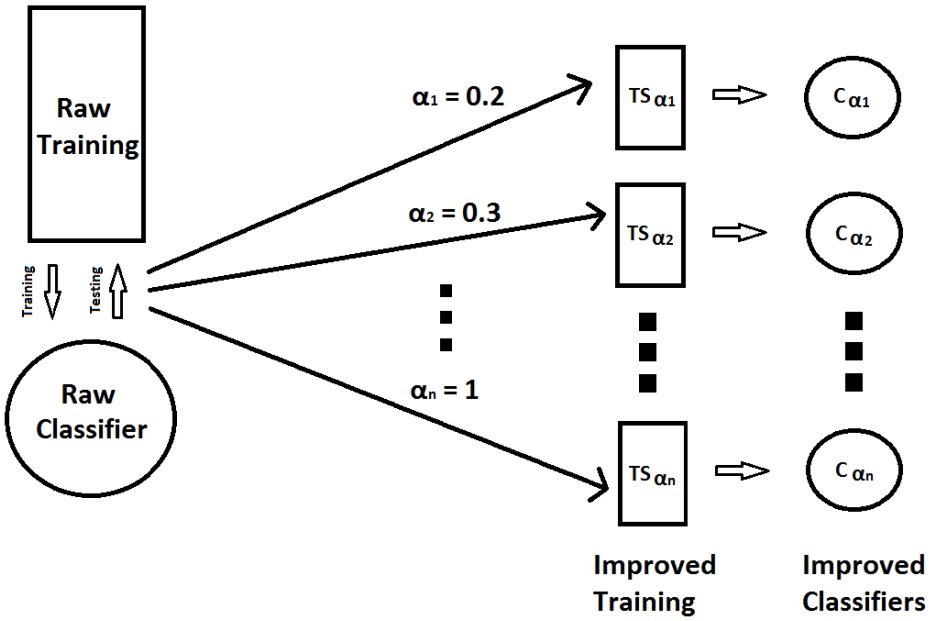


Figure 3.4: Representation of the dataset pruning scheme.

training set.

2. Using the raw classifier to classify all the instances contained in the raw training set. In other words, using the raw training set in place of a test set.

3. Saving the results of this classification to have, for each instance:

   - The corresponding class, for which the instance has been downloaded;
   - The class predicted by the classifier;
   - The **confidence factor** of the classifier in predicting the class.

     The Näive Bayes algorithm classifies a given instance based on the class with the maximum posterior probability distribution given the observation. So this probability is used as a confidence factor in predicting a given class.

4. Removing, from the previously saved data set, the incorrectly classified instances.

5. Obtaining different Training Sets $TS_{\alpha 1}$, $TS_{\alpha 2}$, ..., $TS_{\alpha n}$ by applying a variable threshold α from 0.2 to 1. The threshold is used to remove all the instances that have been classified correctly, but with a confidence factor lower than the threshold value.

It has been decided to produce also another training set, obtained just by removing the incorrectly classified instances, without applying any threshold to the correctly classified instances: this training set is called T_0.

All the resulting training sets have been used to train different classifiers $C_{\alpha 1}, C_{\alpha 2}$, ... , $C_{\alpha n}$, whose parameters have been selected by the optimization process previously described. Table 3.8 shows the parameters, the n-grams and the number of features used for each classifier.

Table 3.8: Parameters optimization results.

Classifier	N-Gram (max)	Features
Raw	2	6760
$\mathbf{C_0}$	2	4800
$\mathbf{C_{0.2}}$	2	4800
$\mathbf{C_{0.3}}$	2	4800
$\mathbf{C_{0.4}}$	2	4800
$\mathbf{C_{0.5}}$	2	4760
$\mathbf{C_{0.6}}$	2	4760
$\mathbf{C_{0.7}}$	2	4840
$\mathbf{C_{0.8}}$	2	4760
$\mathbf{C_{0.9}}$	2	4400
$\mathbf{C_1}$	2	4000

3.2.4 Results

In this subsection, we present the results obtained, on a standard test set, by the different classifiers. Since all the classifiers have been trained in the same way, but with different training sets, it was possible to assess the effectiveness of the data set pruning technique introduced and to evaluate which threshold allows one to obtain the best performance.

3.2.4.1 Test set

The choice of the test set is a critical element, for evaluating the performance of a classifier. In this section, we have derived a test set from the EmoTweet-28 dataset [125]. This dataset consists of tweets manually classified according to 28 different emotions. Since we only need a subset of these emotions, we have defined some classes of EmoTweet-28 emotions, that can be associated with each of the considered primary sentiments. In Table 3.9 this process is summarized:

Table 3.9: EmoTweet-28 classes used as representative of Parrot's primary sentiments. The tweets corresponding to classes of EmoTweet-28 not reported in the table have not been included in the test set.

Macro-categories	EmoTweet-28 Emotions
Joy	"Amusement", "Excitement", "Happiness","Inspiration", "Pride"
Love	"Fascination", "Love"
Surprise	"Surprise"
Anger	"Anger", "Hate", "Jealousy"
Fear	"Fear"
Sadness	"Sadness", "Regret", "Sympathy"
Objective	"none"

Since many of these tweets are labeled with more than one emotion, it has been decided to maintain only the tweets with associated emotions of the same macro-categories according to Table 3.9. Furthermore, the pre-processing stage as described in subsection 3.2.1 has been applied. Finally, as mentioned in the previous chapter,

considered the large amounts of objective tweets, it has been decided to remove some of these from the test set and to insert them in the raw training set.

As a consequence, the test set possesses 10499 instances subdivided for each class as follows:

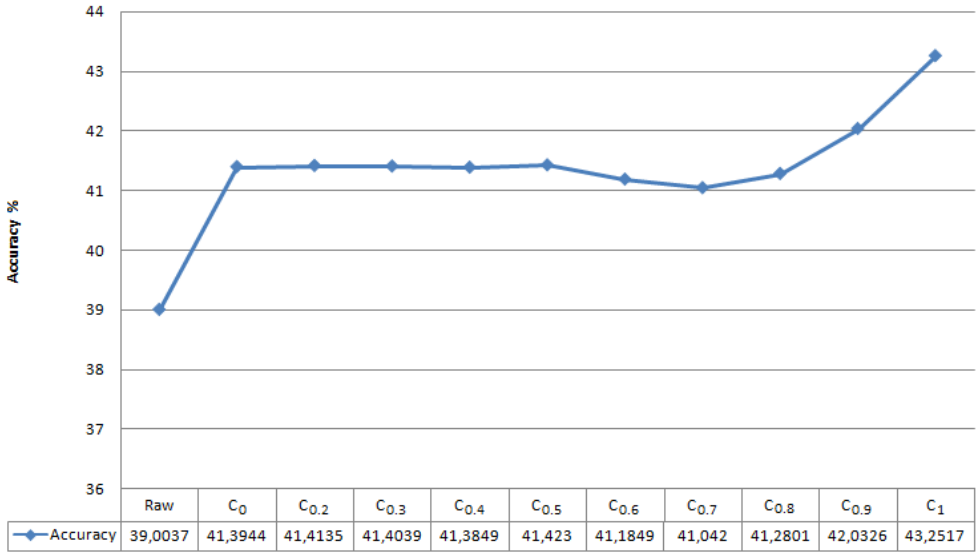
- Joy: 2781;
- Love: 447;
- Surprise: 15;
- Anger: 1221;
- Sadness: 98;
- Fear: 204;
- Objective: 5733

3.2.4.2 Analysis of the accuracy

In this subsection, the results obtained on the test set by the original raw classifier and the improved classifiers are described.

Figure 3.5 shows the accuracy for each classifier. The raw classifier has an accuracy of 39,00% and all the other classifiers, obtained using the different filtered training sets, improve the accuracy to some degrees. More in detail, note that even the classifier C_0, from which only wrongly classified instances have been removed, and no threshold has been applied, allows boosting the accuracy of the results.

To have a better understanding of the performances of the classifiers, we present in Figure 3.6 the F-measures obtained through each classifier. The figure shows that the impact of the data pruning technique is not the same on all classes, possibly because of the different average certainty degree of the different classes, which may cause the filter to alter the balance of the original data set. However, if one considers the average F-measure over the seven classes, a steady increment in the global performance can



	Raw	C_0	$C_{0.2}$	$C_{0.3}$	$C_{0.4}$	$C_{0.5}$	$C_{0.6}$	$C_{0.7}$	$C_{0.8}$	$C_{0.9}$	C_1
Accuracy	39,0037	41,3944	41,4135	41,4039	41,3849	41,423	41,1849	41,042	41,2801	42,0326	43,2517

Figure 3.5: Visualization of the accuracy obtained by the different classifiers on the given test set.

be observed. It should be noticed that this measure is independent of the *a-priori* distribution of the test data among the seven classes.

The aforementioned results lead to two important observations:

- Even if the C_1 classifier produces a small increment of the F-measure concerning the "surprise" class, the low F-measures of the class **Surprise**, obtained by all the classifiers, are probably related to the lack of a suitable number of instances in the test set. EmoTweet-28 contains many tweets associated with the *Surprise* label; however, many of these were ignored since *Surprise* was not the only label assigned to them. This caused very few instances of that class to be included in the test set.

- The improved classifiers could not obtain an increment of the "fear" F-measure. The reason is probably related to the fact that Twitter users are hesitant about

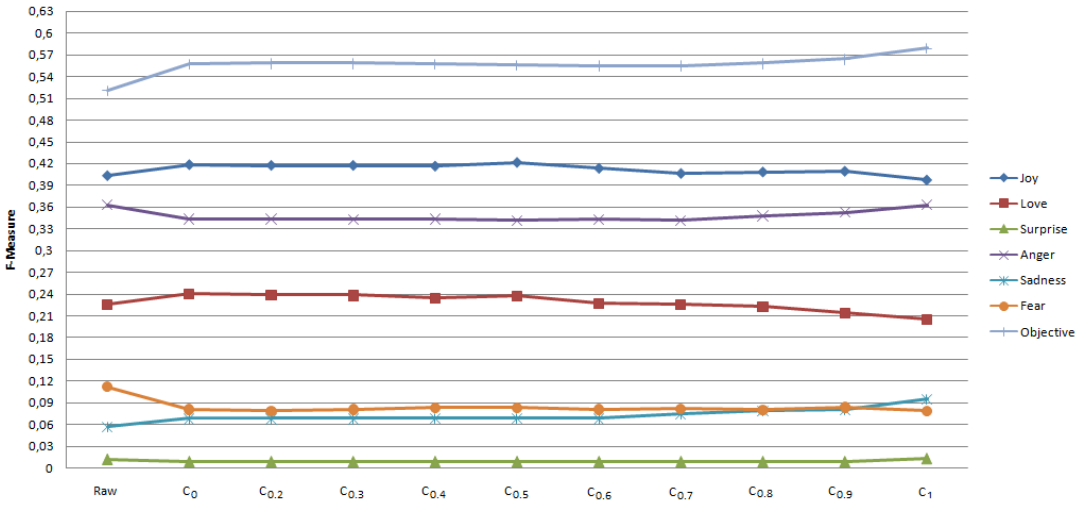


Figure 3.6: F-measure trends for each classifier.

sharing their real fears. It follows that the distant supervision approach is not effective with this type of class. So, the reduction of the trend can be explained by the fact that, for this particular class, the hypothesis of applicability of the considered dataset pruning technique is not verified, since the percentage of spurious instances is superior to the percentage of correct ones.

3.3 Comparison of classifiers for Sentiment Analysis

A common approach to sentiment analysis includes two main classification stages, represented in Figure 3.7:

1. Subdivision of texts according to the principles of objectivity/subjectivity. An objective assertion only shows some truth and facts about the world, while a subjective proposition expresses the author's attitude toward the subject of the discussion.

2. Determination of the polarity of the text. If a text is classified as subjective, it is regarded as expressing feelings of a certain polarity (positive or negative).

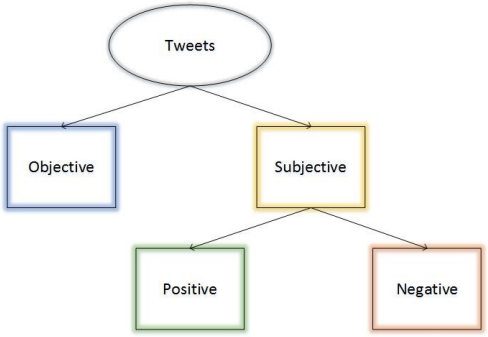


Figure 3.7: Basic classification.

The purpose of this section is to show research aiming at improving the existing basic classification of tweets. Within this context, improving the classification should be considered as an extension of the basic model in the direction of specifying the emotions which characterize subjective tweets, based on Parrott's socio-psychological model. According to it, all human feelings are divided into six major states (three positive and three negative):

- *positive feelings* of love, joy, surprise;
- *negative feelings* of fear, sadness, anger.

In this research, a flat and a hierarchical classifier, which are shown in Figure 3.8 and Figure 3.9, respectively, are considered.
Hierarchical classification is based on the consistent application of multiple classifiers, organized in a tree-like structure. In the case under examination, a first step uses a binary classifier that determines the subjectivity/objectivity of a tweet. The second step further processes all instances that have been identified as subjective. Then, another binary classifier that determines the polarity (positivity/negativity) of a tweet is employed. Depending on the polarity assessed at the previous level, the third step

classifies the specific emotion expressed in the text (love, joy or surprise for positive tweets; fear, sadness or anger for negative tweets).

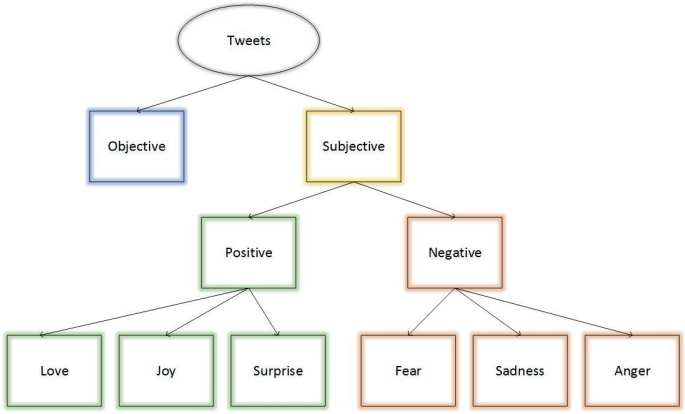


Figure 3.8: Hierarchical classification.

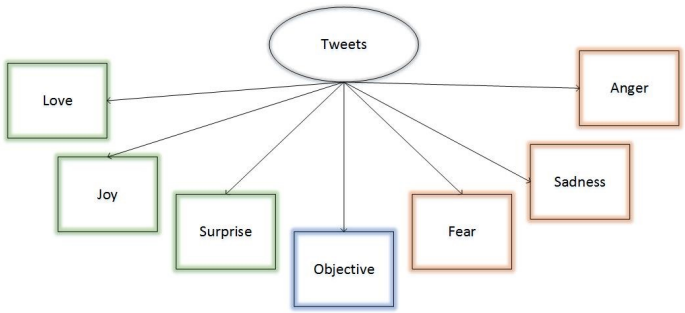


Figure 3.9: Flat classification.

To limit the need for human intervention in the definition of the training data, and thus allow for the collection of larger data sets, a strategy to completely automate the collection of one training set for the construction of the flat multi-classifier, and other four for training the classifiers comprised in the hierarchical model have been devised.

Manual labeling has been used only in building the test sets, since the reliability of such data is critical for the evaluation of the results of the classifiers taken into

consideration.

The rest of this section briefly describes the modules that have been developed to implement the just described method, as well as the data and the procedure adopted to create the training sets.

The research has been developed using Java within the Eclipse IDE. It is structured into three main modules, as shown in Figure 3.10.

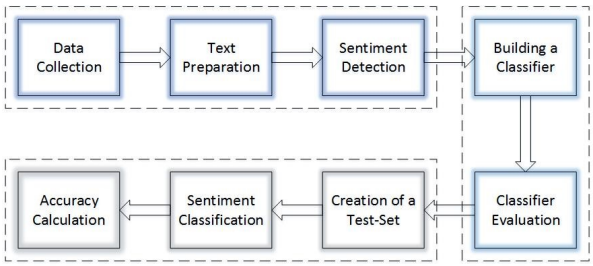


Figure 3.10: Application structure.

3.3.1 Collection of training data

The primary requirement for constructing an emotion classifier based on a machine learning approach is to download a sufficient amount of posts for the training phase. Tweets must be pre-processed, clearing them from the elements which have no emotional meaning, such as hashtags and user references. It is also important to correct spelling mistakes and to encode special characters and emoticons appropriately as text tokens. Each sample of the training set represents a tweet and is composed by the processed text and an emotion class used as a label.

3.3.2 Training sets and classification

The considered classifiers were trained using the “Naive Bayes Multinomial” algorithm provided by Weka. They have been trained using training data collected automatically and systematically into the training sets that contain, in each line, one tweet and the label of the class it belongs to.

For the feature selection, we first used a Weka filter (*StringToWordVector*) to turn a string into a set of attributes representing word occurrences. After that, an optimal set of attributes (N-grams) was selected using the *Information Gain* algorithm provided by Weka, which estimates the importance of a feature by measuring the information gain with respect to the class.

For the hierarchical classifier, four training sets have been created, with data labeled according to the task of each of the four classifiers: "OBJ" or "SUB" for the objectivity classifier, "POS" or "NEG" for the polarity classifier. For the lower-level emotion classifiers (one for the tweets labeled as positive, one for those labeled as negative by the higher-level classifiers) the following labels/classes have been considered:

- Classes of positive polarity: LOVE, JOY, SURPRISE;
- Classes of negative polarity: ANGER, SADNESS, FEAR.

For the flat classifier, a multi-class file was created, with each sample labeled according to one of the seven classes, six representing the considered emotions and the seventh for the objective tweets. Namely: LOVE, JOY, SURPRISE, ANGER, SADNESS, FEAR, OBJ.

The channels used to obtain emotive tweets (according to the the corresponding hash-tag) involve emotions that Parrott identified as either primary, secondary, or tertiary. Parrot's taxonomy of basic human feelings, including only the primary and secondary level, is presented in Table 1.

The training set defining the objectivity or subjectivity of a tweet has been downloaded from the SemEval3 public repository. ([4])

3.3.3 Classifying data

The function library provided by Weka has been used to develop a Java application for assessing the quality of the considered classifiers. The application supports classification models taken into consideration for processing a test set using the Weka

[4]https://www.cs.york.ac.uk/semeval-2013/task2/index.php?id=data.html

Table 3.10: Primary and secondary emotions of Parrott's socio-psychological model.

Primary emotion	Secondary emotion
Love	Affection, Lust/Sexual desire, Longing
Joy	Cheerfulness, Zest, Contentment, Pride, Optimism, Relief
Surprise	Surprise
Anger	Irritability, Exasperation, Rage, Disgust, Envy, Torment
Sadness	Suffering, Sadness, Disappointment, Shame, Neglect
Fear	Horror, Nervousness

classifier models trained with the data described above, labels data and assesses the classifiers' accuracy by comparing the labels assigned to the test data by the classifiers to the actual ones, reported in the test set.

3.3.4 Results of the comparison

In this subsection, the results of the just described comparison between classifiers are presented. First of all, the experimental setup has been described with a particular focus on the procedure followed to collect the data sets for training and testing the classifiers, then the preliminary tests made to evaluate the quality of data and to determine the optimal number of features as well as the size of the N-grams used as features are presented.

Finally, the comparison between the flat and the hierarchical classifiers on the basis of the accuracy they could achieve on the test set is carried out and accurately shown.

3.3.4.1 Collection of the data

The considered sets were built in a completely automated way, without human intervention and according to the following operations:

- *Raw training set (Training Set 1).* The raw training set (in the following called TS1) consists of about 10,000 tweets: about 1500 tweets for each emotion

were collected and as many objective tweets. For the six nuances of emotions, data coming from several Twitter channels were gathered, following Parrott's classifications. Thus, the selection of channels was made methodically, without human evaluation. For each emotion, all the three levels of Parrott's model was used: for example, to extract tweets expressing *sadness* the data from the channel related to the primary emotion, *#Sadness*, but also from those related to secondary (*#Suffering, #Disappointment, #Shame, . . .*) and tertiary emotions (*#Agony, #Anguish, #Hurt* for *Suffering*; *#Dismay, #Displeasure* for *Disappointment*, and so on) have been downloaded. The objective (neutral) tweets were selected from the data set used for the SemEval competition ([5]).

- *Refined training set (Training Set 2).* Since the raw training set contains tweets obtained directly from Twitter channels, it may undoubtedly contain spurious data. Thus, an automatic process to select only the most appropriate tweets has been adopted. TS1 has been filtered to remove the most ambiguous cases, and a second training set (in the following called TS2) of about 1000 tweets for each of the six primary emotions has been obtained. The filtering process was based on six binary classifiers, one for each emotion. The training set for each of them was balanced and considered two classes: the "positive" class included all raw tweets automatically downloaded from sources related to the emotion associated with the classifier; the "negative" class included tweets coming, in equal parts, from the other five emotions and the set of objective tweets. Finally, TS2 included only the tweets which could be classified correctly by the binary classifier, for the tweets used for training the main classifiers (i.e., those in TS2) to be as prototypical as possible.

- *Test Set.* Tweets for the test set were downloaded in the same way as those for the training set, but they were manually annotated. They consist of 700 tweets, 100 for each of the six emotions in addition to 100 objective tweets. Even if a representation sufficiently relevant for a correct classification would require a much larger number of features, their first two components obtained

[5]http://en.wikipedia.org/wiki/SemEval

by Principal Component Analysis (PCA) have been plotted in Figure 3.11 to give a first rough idea of the distribution of the tweets in the feature space. Objective tweets (yellow) and tweets related with sadness (green) are quite clearly separated from the others even in this minimal representation. Instead, other emotions are much closer and significantly overlapped, especially those related with surprise (violet). This could actually be justified considering that secondary and tertiary emotions can play a very significant role in recognizing this emotion since it can be equally associated with both positive and negative events.

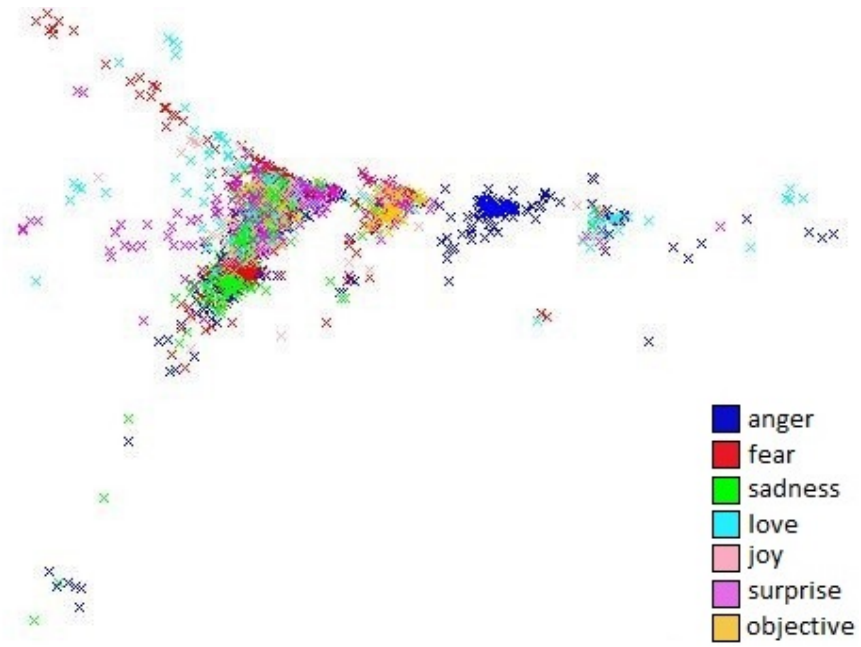


Figure 3.11: Visualization of the first two PCA components for the test set.

3.3.4.2 Optimization of the parameters of the classifiers

For each classifier (four for the hierarchical and one for the flat approach), a systematic preliminary analysis was performed to optimize some relevant parameters that affect

the training phase. A grid of configurations was selected, and then cross-validation was used to estimate the quality of classifiers configured according to it. In particular, the optimal length of N-grams to be used as features was researched. Figure 3.12 shows the case of the flat classifier, but the other cases are similar. It can be observed that accuracy nearly peaks at N-gram = 2. Longer sequences increase the complexity of the training phase, without producing any significant improvement of the results. The dependence of the performance on the number of features selected has also been analyzed using Weka's Information Gain algorithm. In Figure 3.12 one can observe that its increase does not provide a monotonic improvement of the classifier quality. Instead, it has a peak at around 1500 features.

Table 3.11 shows the results of the parameter optimization step. In particular, the N-Gram (max) value is 2 for all the considered classifiers (unigram and bigram are considered). The last column shows the number of features that optimizes the performance of the classifiers.

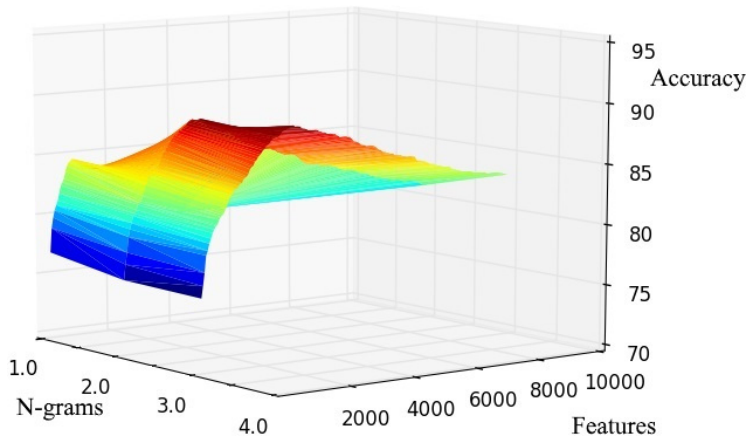


Figure 3.12: Optimization of the accuracy vs. N-grams and number of features.

Table 3.11: Parameter optimization results.

Classifier	N-Gram (max)	Features
Flat	2	1500
Sub/Obj	2	2500
Pos/Neg	2	2500
Anger/Sadness/Fear	2	1550
Love/Joy/Surprise	2	1500

3.3.4.3 Analysis of the accuracy

Tables 3.12 and 3.13 report the results obtained on the test set by the seven-output flat classifier and by the hierarchical classifier.

Table 3.12: Accuracy of flat classifier, using alternatively the two training sets.

	Flat					
	TS1			**TS2**		
	Prec	**Rec**	**F_M**	**Prec**	**Rec**	**F_M**
Objective	**0.70**	0.69	0.69	0.58	0.73	0.65
Anger	**0.36**	0.18	0.24	**0.36**	0.23	0.28
Fear	0.39	0.22	0.28	**0.50**	0.13	0.21
Sadness	0.29	**0.68**	0.41	0.31	0.66	0.42
Love	0.45	0.29	0.35	**0.50**	0.32	**0.39**
Joy	0.32	0.34	0.33	0.36	**0.42**	**0.39**
Surprise	0.46	0.39	0.42	**0.50**	0.42	0.46
Total accuracy	40,48%			42,33%		

They show the accuracy of each approach when trained on TS1 or on TS2, to assess the effect of the refinement step. These tables confirm the advantage of a hierarchical classification, which intrinsically exploits the *a priori* domain knowledge embedded into the whole classifier structure. They also show that some emotions (e.g., sadness) are classified rather well, while others are harder to classify (e.g., anger).

For each class, the best results regarding precision, recall, and F-measure have been emphasized. These results show that the best results have been obtained using the filtered training set (TS2).

Table 3.13: Accuracy of hierarchical classifier, using alternatively the two training sets.

	Hierarchical					
	TS1			**TS2**		
	Prec	**Rec**	**F_M**	**Prec**	**Rec**	**F_M**
Objective	0.66	0.87	**0.75**	0.60	**0.88**	0.71
Anger	0.31	0.27	0.28	0.34	**0.30**	**0.32**
Fear	0.40	0.22	0.28	0.46	**0.23**	**0.31**
Sadness	0.34	0.56	0.43	**0.38**	0.60	**0.46**
Love	0.40	**0.34**	0.37	0.43	0.28	0.34
Joy	**0.39**	0.34	0.37	0.39	0.38	0.39
Surprise	0.47	0.41	0.44	0.50	**0.45**	**0.48**
Total accuracy	43,61%			**45,17%**		

Table 3.14 reports in detail the partial results of the three classification levels of the hierarchical classifiers: the first and the second level of classification, i.e., subjectivity and polarity, have an accuracy of around 90% and 75%, respectively. Aggregating the results of the flat classifier to provide the same partial responses provides systematically worse results. This is not surprising, since a seven-output

Table 3.14: Accuracy of the intermediate results of hierarchical classification, based on TS1 and TS2.

	TS1	TS2
Objective / Subjective	91.90%	90.06%
Positive / Negative	73,36%	75,54%
Final classification	43.61%	45.17%

classification is a more laborious task in general, and for ambiguous (and often mixed) emotions in particular. On the other hand, the cascaded structure of a hierarchical classifier has a higher risk of propagating errors from the higher levels to the lower ones. From this point of view, the results show that the structure we adopted minimizes that effect, since, still not surprisingly, the accuracy of the classifiers increases with their level in the hierarchy.

Finally, Table 3.15 shows the confusion matrix of the hierarchical classifier trained using TS2, which is the best performing approach in the considered research. Notably, fear and anger are often misclassified as sadness and love is often misclassified as joy or surprise.

Table 3.15: Confusion matrix of the hierarchical classification based on TS2.

–>	Objective	Fear	Anger	Sadness	Love	Joy	Surprise
Objective	**88**	3	2	3	2	1	1
Fear	15	**22**	15	23	1	13	5
Anger	15	7	**30**	38	4	5	1
Sadness	8	1	13	**63**	1	10	8
Love	5	8	8	11	**29**	20	21
Joy	8	4	7	16	17	**39**	10
Surprise	7	2	12	12	12	11	**47**

Part II

Sentiment Analysis Applied to Social Networks

Chapter 4

Results of the application of Sentiment Analysis to Social Network

In this part of the thesis it will be investigated how sentiment analysis techniques, together with a combination of automatic learning, natural language elaboration, network analysis and statistics, can result fruitful to measure the human behavior in the so-called *social media analytics*, which in turn can result effective in various societal fields such as, for example, in penal forensics investigations, marketing strategies [41], and so on.

4.1 A combined approach of Sentiment Analysis and Social Network Analysis

This section describes the results obtained by combining two approaches, namely sentiment analysis (SA) and social network analysis (SNA), to analyze communities of users on Twitter. In particular, a sentiment has been associated with the nodes of the social network showing social connections, and this helps to highlight the potential

correlations. The idea behind it is that, on the one hand, the social network topology can contextualize and then, in part, unmask some incorrect results of the sentiment analysis; on the other hand, the polarity of the feelings on the network can highlight the role of semantic connections in the hierarchy of the communities that are present in the social network itself.

4.1.1 Motivation

Twitter is a platform which may contain opinions, thoughts, facts, references to images and other media and, recently, stream video filmed live and uploaded by users. So it is more than just a social network where a user shows and increases his social relationships; it is a real communication channel in which a user can choose his topics and his nodes of reference according to his interests and culture. A study of the network topology and the number of interconnections of a node are able to highlight the communities in the network and also, in part, how the information is propagated, but they are not able to say anything about the degree of agreement and cohesion of the members of a community. To solve this task, an investigation into the semantic content of the messages should be carried out. This could be done through sentiment analysis, which, however, has some difficulties in terms of effectiveness. This is mainly due to the subtle distinction that exists between positive and negative sentiments or between neutral and positive one. For example, a sentence containing irony or sarcasm, where the interpretation of the meaning is strictly subjective, would make two human beings to be in disagreement about the real feeling that it expresses. Furthermore, not always the opinions are expressed through the use of opinion words, in many cases special language constructs (such as the figures of speech) come into play. Difficulties arise also from the use of non-formal expressions and slang that do not belong to the vocabulary of a language. These terms are often used intensively to express a particular opinion or a certain mood. Additional problems are due to the domain of the subject: in particular, it can be noted that the feelings that are expressed by a word are often dependent on the topic.

As a microblogging service, Twitter is used to publish short messages counting a maximum of 140 characters (tweets). This characteristic on one side forces people

to take a position, on the other side the few words do not allow the user to repeat concepts or emotions: he rather uses slang shared by the community, emoticons, and punctuation. Despite the ease of retweeting, the difficulty in perceiving what is the real feeling of the user increases and the intense use of citations can also distort the sentiment enclosed in the tweet.

However, by combining the information of SA with that of SNA, one can guess to disambiguate some actual cases and the opportunity to know the slang of the channel under examination can improve the efficiency of machine learning algorithms for the SA.

4.1.2 Social Network Analysis of Twitter: data selection

As a social networking platform, Twitter is structured as a directed graph, in which each user can choose to follow a number of other users (followees), and can be similarly followed by other users (followers). Thus, the *follow* relationship is asymmetrical, it does not require mandatory acknowledgment, and it is essentially used to receive all public messages published by any followee user. Consequently, in the considered analysis, three types of data (Fig. 4.1) have been collected: the User type that represents users' profiles; the Tweet type that represents posted messages; the Friend type that represents the *follow* relationships among users. Apart from the data obtained directly

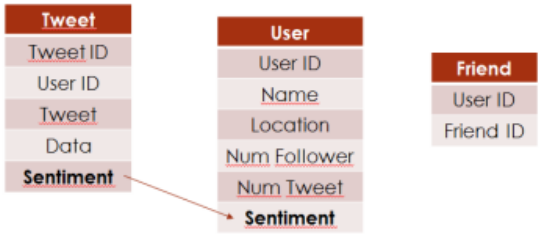


Figure 4.1: Structure of the considered dataset from Twitter.

from Twitter, a field was added to both tweets and users, to associate a sentiment with them, according to the result of the performed SA. If a user has posted more

than one tweet on the network, the sentiment of the last tweet that he posted has been considered.

4.1.3 Text polishing and Sentiment Analysis

As a communication medium, tweets have a quite peculiar nature. Some distinguishing features of the communication on Twitter are related to technical aspects: those including the length of text, tags, URLs, etc. Other features may be classified as idiomatic use of the medium, and create a sort of Twitter culture. As a starting point, a tweet may contain many elements that are not significant for the considered classification, and can thus be dropped through a filtering process. To polish the message, various filters can be applied in a customizable sequence. An example is shown in Fig. 4.2.

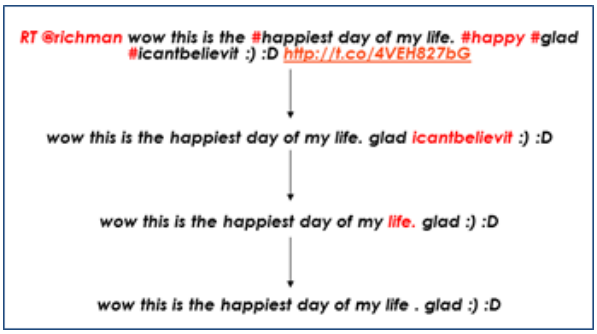


Figure 4.2: Sequence of operations to clean a tweet.

A first filter eliminates useless tokens such as: the *RT* sequence; the @ character and the whole following username; the # symbol, but not the following topic name, which is kept in the message. The topic name is also removed when it coincides with the name of the channel where tweets are collected from.

A second filter applies the language-specific rules. It includes an orthographic correction of the message, which is used to remove unknown words (in the example: "icantbelievit") and other filtering processes for stemming and removal of stop-words.

Finally, another filter separates all punctuation symbols from the text and organizes

them as single-character words. Even if smiles sequences, repeated question and exclamation marks are kept as aggregates because they are essential patterns for the classification. The final result of the filtering process is a word vector, which is then submitted to a set of classifiers.

A set of classifier has been used to identify the following classes of messages: undiscriminated, objective, subjective, positive, and negative. Moreover, there is a class in which the system insert all the tweets that are too short to be classified. The system is organized as a simple hierarchy of agents, mimicking the hierarchy of sentiment classes. Since objective messages have no polarity by definition, the classifier for positive and negative sentiments is only applied to subjective messages (see Fig. 4.3). One advantage of this framework for classifiers is the ease with which one can add classifiers trained to identify other emotions. In fact, hierarchical classification has been applied successfully in a number of studies, for information retrieval [115] and it has been proven effective especially in the case of classification over hierarchical taxonomies. Also in the case of sentiment analysis, a hierarchy of classes can be defined [50, 12]. Accordingly, hierarchical classification has already been applied to sentiment analysis, too [115]. Each classifier is based on Multinomial Naive Bayes

Figure 4.3: Hierarchy of basic sentiment classes.

algorithm, one of the most popular methods used in SA. It has been selected because it seems to be the most suitable to generate and process large sets of features. In fact,

instead of generating a training set by hand, it has been realized an automated (or at least semiautomated) process for obtaining good training sets. In the considered methodology, the training sets are obtained through the automatic elaboration of some particular streams of tweets and comments, obtained directly from Twitter, without any manual classification. Thus, each training set may contain an important number of wrong data. Nevertheless, they can be used to obtain useful results.

As concerns, the objectivity/subjectivity classifier, a strategy similar to the one used in [101] has been used. In fact, to obtain objective content, messages generated from popular news agencies have been gathered.

In the considered tests, the following list has been used: @ABC, @BBCNews, @BBCSport, @business, @BW, @cnnbrk, @CNNMoney, @fox32news, @latimes, @nytimes, @TIME. To obtain subjective content, instead, comments directed to the same list of users have been gathered.

As regards the polarity classifier, different sources have been used, thus generating training sets which do not overlap with those about objectivity/subjectivity. Sources of mostly positive or negative messages have been used respectively. On the one hand, those sources should fit the particular setting of Twitter (short messages, idiomatic expressions, smiles, etc.). On the other hand, they should be not specific to a particular topic or context (sport, music, etc.). Thus, the idea of collecting messages about particular events, mostly generating either positive or negative sentiments, has been dropped. Instead, messages have been collected using generic yet polar terms as queried hashtags. In particular, the following channels have been used to gather positive content: #adorable, #awesome, #beautiful, #beauty, #cool, #excellent, #great. Conversely, the following channels have been used to gather negative content: #angry, #awful, #bad, #corrupt, #pathetic, #sadness, #shame. Actually, such terms have been chosen quite empirically, taking into account the quality of the training sets they generated. However, they could be selected from WordNet-Affect [118], SentiWordNet [10], and other effective lexicons, in a more systematic way. In this way, the training set is generated in an automated fashion, as a list of tweets. Each tweet is associated with its supposed class, in accordance to its source. In fact, the training set is not perfect, as it contains messages gathered from public channels. However, a training

set of this kind can be generated easily and in a methodical way, from real and updated Twitter messages. Moreover, it is possible to extend this approach to train a classifier in order to recognize feelings which are written in a particular slang.

Feature	$P(F_i \mid pos)$	$P(F_i \mid neg)$
:)	**0,0025**	0,00055
stupid	0,000098	**0,00065**
thank you	**0,0012**	0,00029
!!!	0,0028	0,0019

Feature	$P(F_i \mid obj)$	$P(F_i \mid sub)$
!!!	0,000031	**0,002**
:(	0,0000079	**0,00030**
%	**0,0013**	0,00080

Figure 4.4: Generated model for the considered classifier: example of selected features and their probabilities in the polarity (on the left) and subjective (on the right) classifiers.

In Fig. 4.4 some examples of features which are selected by the classifiers together with their probabilities are presented. It is worth noting that these are consistent with what was expected: the emoticons ':)' have a high probability of being in positive phrases, while the pattern '!!!' is very significant for the classifier of subjectivity, but it is a useless feature to determine the polarity of a tweet.

4.1.4 Experimental Results

In this subsection, the results of the classifiers and the analysis carried out on a couple of case studies are shown. Using the methodology which has been previously described, it is possible to obtain some generic training sets for the classifiers. This phase was carried out before selecting the final case studies. In the considered settings, they consist of:

- 86000 instances (polarity);
- 32000 instances (subjectivity).

These instances have been obtained by exploring more than 60 channels on the social network. In the generated models, the selected features are consistent with some possible expectations: the typical expressions of a certain feeling (such as smileys, or some words that express appreciation or disgust) show a higher probability of

belonging to the class of that feeling, rather than to the class of the opposite sentiment. The results obtained by the classifiers using cross-validation (with folds = 10) on the training sets showed an accuracy of:

- 77,45% (polarity classifier)
- 79,50% (subjectivity classifier)

These results show that the model of the classifiers contains effective features for the recognition of the sentiment of a message. The case study which was considered in these experiments is the social network of the #SamSmith channel (the singer who won four awards at the Grammy Awards 2015). The choice of this channel is justified by the strong similarities found between the type of the published tweets and the instances used to train the classifiers. All data were downloaded between 2015-02-02 and 2015-02-10. The awarding of the Grammy took place on 2015-02-08. The social network (shown in Fig. 4.5) consists of a total of 5570 nodes (users) and 6886 arcs (*follows* relationships). Nodes are deployed according to the ForceAtlas2 algorithm [66], which turns structural proximities into visual proximities, thus highlighting communities. As

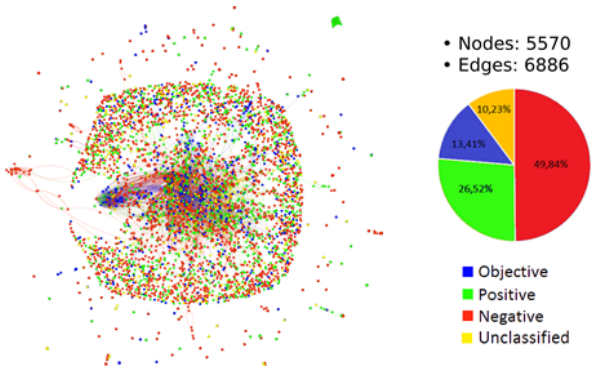


Figure 4.5: Combined analysis of the #SamSmith channel.

shown in Fig. 4.5, the prevailing sentiment detected from the classifier is the negative one. Performing an analysis on a sample of tweets in the network, it can be noticed that many sentences are actually quotes of songs. These messages contain melancholic

and sad phrases, and are therefore classified as negative. Considering that a quote is generally an appreciation for the artist, most users classified as negative are actually positive users. This is a typical example of a classic problem of misunderstanding of SA: the system, while classifying correctly the tweet, misses the assessment of the feeling because it can not evaluate the tweet together with its context.

In order to evaluate the performances of the system, a simple survey through a group of people has been conducted. In this way, 100 messages that show a clear opinion on the singer have been selected and classified. Then, those messages have been used as a test. The results of the classifiers showed an accuracy of 84% for the polarity and 88% for subjectivity.

In the network periphery (at the top-right corner of Fig. 4.5), it is possible to notice a small group of users whose feeling is completely positive. After a careful analysis of users' tweets in this small group, it was found that these posts are mainly retweets and the original messages are only two. Of these two messages, the first is actually positive, while the other one is objective. This episode shows how some errors of assessment can have important impact on larger communities. In addition to the #Samsmith channel, the social network associated with the #Ukraine channel has been considered, trying to obtain some particularly significant results, above all from the point of view of network topology. In fact, the crisis in the region could have led to a quite sharp division on the Web. The network consisted of:

- 26131 nodes
- 1163588 edges

In Fig. 4.6, it is possible to see the main results of the analysis on the network. The more evident thing to notice, is that the prevailing color in the network is blue (objective tweets), and the next one is red (negative tweets). Given the nature of the channel under consideration, which essentially reflects a social tragedy, the sentiment found through the analysis is quite plausible. However, analyzing some random messages, it can be noticed a number of errors in the classification of these tweets. In particular, some objective sentences are often classified as negative ones, while some sentences expressing essentially hope (and thus positive) are classified as objective

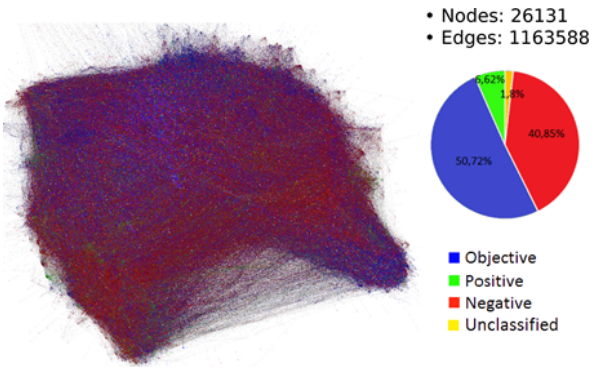


Figure 4.6: Combined analysis of the #Ukraine channel.

ones. The reason for these errors could be related to the type of features contained in the model of the classifiers, which could not be a good fit for this particular case study. The case of Ukraine has been discussed quite largely in traditional media, too, for the supposed role of "trolls" operating on new media to influence the public opinion. In fact, this may represent, as a modern reproposition, the quite classical case of opposing propaganda campaigns, this time carried on through social media. Also for this reason, the social communities participating in the channel have been analyzed. The focus has been on the most active users, who contributed with at least 6 tweets during the whole week that has been considered (mid July 2014). In fact, among those it is more probable to find candidate opinion makers. The analyzed subnetwork represents around a tenth of the original network, and precisely consists of:

- 3261 nodes
- 84307 edges

The community detection algorithm provided with Gephi, at various resolution levels, has been used [81]. Quite interestingly, it has been possible to identify quite clearly two major communities. Additionally, some much smaller communities were found.

Looking at data reported in Table 4.1.4, it is easy to notice that the two communities, corresponding to opposing factions in the crisis, have a quite similar size.

	Full Network	**Community 1**	**Community 2**
Average degree	51.706	53.021	42.649
Diameter	7	7	6
Radius	1	4	4
Avg path length	2.511	2.248	2.334
Shortest paths	10591776	3152400	2014980
Graph density	0.016	0.030	0.030
Clustering coeff.	0.420	0.480	0.414
Total triangles	873460	540526	281524

Table 4.1: Features of the main communities detected on the #Ukraine channel.

Moreover, also their main features are quite similar. This seems to indicate that the two campaigns have a quite similar internal social organization, at least at the macroscopic level. Nevertheless, both the communities have high density, almost doubling the value of the whole network. This means that there is a quite clear separation between those two communities, which have relatively few shared connections. The considered sentiment analysis has not highlighted significant differences in the emerging opinions in the two communities. In fact, they largely share the same negative outlook of the whole network. This is an issue that could be analyzed in deeper detail in the future. The emerging sentiment in each campaign may also vary during time, and in particular in correspondence with major events and turn-points in the crisis.

4.2 A Case Study for Emotion Detection

This section analyzes the online debates about “Brexit” before, around and after the referendum for reaming or leaving the European Union held in the United Kingdom and Gibraltar on June 23^{rd}, 2016. The online debate generated a lot of textual data carrying emotional information and in this section the results obtained by applying emotion classification to tweets about Brexit referendum are presented. The classification approach is based on a three-level hierarchy of four specialized classifiers,

which reflect relationships between the target emotions.
The analyzed emotions are those expressed before and after the referendum, studying the impact of Brexit on public opinion. The collected tweets have been geolocalized, in order to analyze feelings and reactions expressed in different UK counties.

4.2.1 Data Collection

The first step was to collect Tweets from the #Brexit channel during the referendum period. Through web scraping from #Brexit it was possible to go back to Tweets ID, the unique identifier for the Tweets, and to use Tweepy Python library to access Twitter API. Tweet ID is the key that allows one to obtain all the information about a specific "status update". Twitter APIs provide those data encoded using JavaScript Object Notation (JSON), however, the APIs work through authentication, with a timeout that limits the instantaneous acquisition. The Tweet Object has a long list of attributes[1], but in the considered case study only some of these were used.

Table 4.2: Used Tweet Object attributes.

Attribute	**Type**
id	Unique identifier for the Tweet.
created_at	UTC time of Tweet creation.
text	The actual UTF-8 text of the Tweet.
user	The user who posted the Tweet.
coordinates	Latitude and longitude of the Tweet.
place	The place indicated by the user.

Tweet Objects are also the "parent" object of several child objects, as in the case of User Object. Each user can be uniquely identified by a user ID and can optionally define a location for his profile and/or enable the possibility of geotagging its Tweets. The goal of this research is to use only geolocated Tweets in the UK so, if a user decides to enable the geotag option on Twitter, it is possible to take advantage of two kinds of real

[1] `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`

location: coordinates and place. Coordinates provides the exact latitude and longitude of the user, for example when a user activate location on a smartphone. A user can also indicate the place he is twitting from, selecting it on a list in his user-interface. APIs provide then a series of longitude and latitude points, defining a bounding box which will contain the place. If it was defined, the exact location was directly used, otherwise the central point of the bounding box was used. From previous studies, however, it can be seen that the number of geolocalized tweets is only about 5% of the total. To increase the number of collected Tweets also the user-defined location for his account's profile (when non-null), expressed as a string of words, was considered. In this way, Google APIs could be used to pseudo-locate Tweets, extracting latitude and longitude from a city name. This method is not really accurate because a user may provide false locations, but this error is negligible. About 570,000 worldwide geolocated Tweets, corresponding to 360,000 users, were retrieved since June 21^{st}, 2016 at 6:00 pm to June 25^{th}, 2016 at 1:00 pm. These Tweets were subsequently filtered by selecting only those from UK with QGIS, a geographic information system (GIS) application that allows management and analysis of geospatial data. This approach brings the amount of users to about 56,000, but they have been further divided into two timelines, *before* and *after* the release of the official referendum results. In this way, it is possible to make a comparison between them and identify the preliminary results. Twitter does not offer any API to provide data about gender, so the gender was retrieved from the name of the users.

Table 4.3: Gender differences.

	Before Brexit	**After Brexit**
Male	74.5 %	66.64 %
Female	25.5 %	33.36 %

Table 4.4: Number of tweets per user.

Before Brexit	After Brexit
1.75	1.54

4.2.2 Preprocessing

This subsection describes the preprocessing techniques used to improve the accuracy of the emotion detection system. A framework that performs different preprocessing Python modules has been employed. The first module manages basic cleaning operations removing URLs, hashtags, mentions and disturbing elements for the next phase of elaboration with Weka, such as quotation marks. All misspelled words are normalized and the punctuation is removed, except for apexes because they are part of grammar constructs, furthermore every Tweet text is converted to lower case. Considering a classical list of emoticons, it was possible to represent them as tags (i.e. :) → smile_happy). This operation is useful for the next module that reduces the number of emoticons to only two categories: smile_positive and smile_negative. During the execution of this module and the following ones the text was made more uniform in order to help the classification in terms of feature selection. One of these modules, for example, replace all negative constructs with "not" and another one applies stemming techniques. Finally stopwords, like pronouns or articles, are filtered to increase classification accuracy.

In addition to positive and negative sentiments, also specific emotions were assigned to each instance (post or comment). In particular, Parrott's socio-psychological model, which classifies all human feelings into six major categories, was considered:

- Three positive feelings: love, joy and surprise;
- Three negative feelings: fear, sadness and anger.

Instances not expressing any particular feeling are taken into account as objective instances.

In order to carry out such a task a *hierarchical classifier*, based on the consistent application of multiple classifiers and organized in a tree-like structure was employed.

In particular, the considered classifier firstly determines the subjectivity/objectivity of an instance, and then further processes each subjective instance, associating it with a polarity; in other words, subjective posts are divided into positive and negative posts. Positive and negative instances are then classified by two separate classifiers that assign them a specific emotion from Parrott's model. This hierarchical classifier is therefore based on a three-level hierarchy of four distinct classifiers, using the *Naive Bayes Multinomial* algorithm.

4.2.3 Training data

Training effectively a classifier destined to carry out such a difficult task requires a sufficiently large training set. Putting together such a set manually would require a significant amount of time. For this reason, in order to limit as much as possible the human interaction, a completely automated approach, known as *Distant Supervision*, has been used.

An initial (or *raw*) training set has been built with data collected directly from Twitter. In particular, around 10,000 tweets have been downloaded and assigned to a class, depending on its hashtags (*i.e.:* a tweet containing hashtags like *#Suffering*, *#Disappointment* or *#Shame* is assigned to *sadness*). These tweets have been pre-processed, in order to clear the elements without emotional meaning, correct spelling mistakes and encode special characters and emoticons as appropriate, before deriving a bag-of-words model, with values calculated using the *TF-IDF* weight function. These attributes have then been filtered using the *Information Gain* algorithm, in order to extract the most meaningful words.

Since data have been collected directly from Twitter, it is extremely probable that they contain some forms of noise. Thus, an automatic process has been employed in order to select only the most appropriate data. A bayesian classifier with seven classes, one for each emotion, has been trained and then tested on the entirety of the training set. Only the instances classified correctly during the testing phase have been used to build a more *refined* training set (Figure 4.7). This refined data set has been finally

used to train the hierachical classifier, then employed to classify the Tweets.

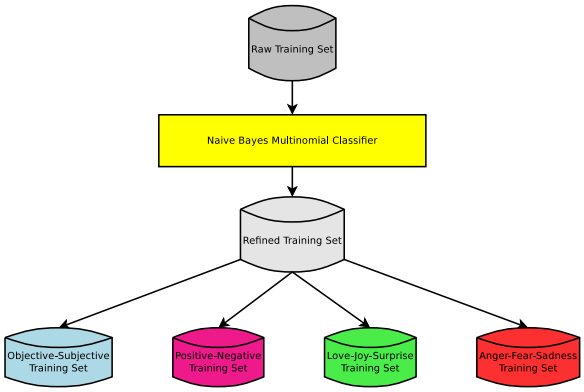


Figure 4.7: Training set structure

4.2.4 Classification

As mentioned, the *Naive Bayes Multinomial* classifier used to classify the posts has been trained using the *refined* data set, described above. Since the considered classifier is hierarchical, in particular, the refined set has been used to create four different training sets, in which data are labeled according to the task of each classifier:

- For the *Objectivity/Subjectivity* classifier, the training set used is simply the refined training set, in which every post associated with a sentiment has been labeled as *subjective*;
- For the *Polarity* classifier, only the subjective instances of the refined training set are considered, divided into positive and negative;
- For the *Positivity* classifier, only the positive instances of the refined training set have been used, divided into love, joy and surprise;
- Similarly, for the *Negativity* classifier, only the negative instances of the refined training set have been used, divided into fear, sadness and anger.

Posts have been pre-processed like the tweets used to create the training set.

4.2.5 Results

As previously mentioned, the tweets regarding Brexit have been divided into two timelines, *before* and *after* the release of the official referendum results, since June 21^{st}, 2016 at 6:00 pm to June 23^{rd}, 2016 at 11:59 pm and from June 24^{th}, 2016 at 00:00 am to June 25^{th}, 2016 at 1:00 pm. In this way, it is possible to make a comparison and check whether there are changes in the obtained data. It is important to note that in the case of multiple tweets, a Python script was developed to generate a GDF file solving the problem. Thanks to the script it was possible to properly represent the user thought a single emotion. It's also interesting to underline that geolocated UK users that tweeted after June 24^{th}, 2016 at 00:00 am number thirteen times higher (51,927) than those that tweeted before (3,867). Studying emotions only, one can see some very interesting results, that deserve to be reported as maps of emotions. Furthermore, it was also checked whether the results of the referendum were predictable by means of the polarity obtained from tweets.

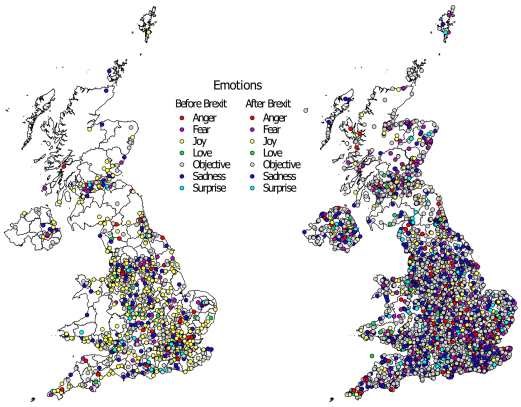


Figure 4.8: Emotions and objectivity map.

Fig. 4.8 represents global emotions, also considering objectivity. However, the feelings caused by Brexit before and after the referendum were analyzed, so objectivity in the considered case is not relevant and it will not be considered any more.

As one can guess from Fig. 4.9, the most striking data are certainly the decrease

of joy by almost 30.8%, but it is fascinating to notice that after the referendum sadness and fear are consistently increased, respectively by 12.5% and by 11%. As a consequence, it was decided to represent the emotions that have changed the most.

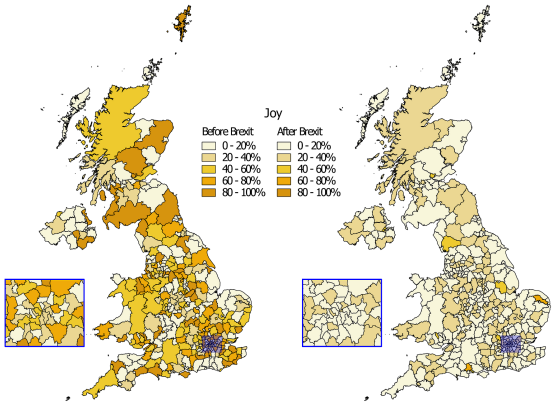


Figure 4.9: Joy map.

It was also possible to represent polarity concluding that in principle the results confirm the initial expectations, with an increase of negative feelings at the expense of the positive ones.

4.3 Analysis of groups in Facebook

In this section a research aiming to analyze the dynamical behavior of the patients of Inversa Onlus Facebook group is presented. The analysis has been performed on the basis of the patients' interactions and emotions expressed in their posts and comments. In particular, the temporal patterns and relationships between the analyzed data are considered. Results have been obtained using a multi-methodology approach, combining:

- **Emotion Analysis**: The emotions expressed in about 50,000 posts and comments written by 1,200 users in the period between 2009 and 2017 have been analyzed using a hierarchical classifier based on Parrot's emotion categorization.
  In order to obtain an accurate system for emotion detection stopwords and stemming algorithms have been used to preprocess the data, the bag-of-words model has been used to extract features from the sentences, the Information Gain algorithm has been used to reduce the features, and the Naive Bayes Multinomial classifier has been used to optimize some hyperparameters in order to achieve better accuracy [25]. In fact, the whole system is built after a selection of the best mechanisms and parameters, in accordance to the state of the art [6].
  Once the classification step has been completed, a further data analysis has been performed considering the resulting emotion information joined with other social media information. In particular, date and time of each post (or comment) and the logic relationships between each post and the comments related to that post have been analyzed.

- **Social Network Analysis**: The results of the Emotion Detection have been joined also with other information retrieved performing a Social Network Analysis. In particular users' activity has been analyzed in terms of:

  1. **Interaction network**, which is a directed and weighted graph where each node represents a user and each link represents the number of comments that have been written by the source node user as a response to posts and

comments written by the destination node user.

2. **Social network**, which is an undirected graph where each node represents a user and each undirected link represents the mutual Facebook friendship between two users (nodes).

4.3.1 Data collection and Privacy preserving

In collaboration with Inversa Onlus association, that manages the support group, it was possible to collect data from the Facebook group using the administration permissions in order to use the official Facebook API. In that way, it was possible to retrieve all the metadata related to posts and comments, including content, author, date and time of publication. About 50000 elements (posts and comments) related to the period between July 2009 and December 2017 have been collected. All the information has been stored only for the agreed analysis and it has been carefully anonymized to preserve the privacy of the members.

4.3.2 Creation of the training set

The creation of an adequate training set for the emotion analysis is crucial to increase the accuracy in the classification of patients' emotions. Due to the lack of useful datasets containing annotated posts of Italian patients, it was decided to create a novel training set based on supervised learning and manual annotations. A balanced training set has been built by randomly annotating 10% of all the available content published in the Facebook group in the past seven years (Table 4.5). The main challenge of this manual approach, due to the particular context represented by a patients community, has been the selection of contents conveying positive feelings. *Love* and *Joy* contents have been selected according to secondary and tertiary emotion levels provided by Parrot's categorization [102]. In the light of this, *Love* has been treated as Affection, Caring and Compassion, while *Joy* as Cheerfulness, Optimism and Relief. The unbalanced value of the surprise elements in the training set is motivated by the lack of this kind of positive emotion in the group.

Sentiment	Number of elements in the training set
Objective	400
Love	175
Joy	175
Surprise	50
Fear	135
Sadness	135
Anger	130

Table 4.5: Number of published content used in the training set for each emotion.

4.3.3 Pre-processing and features selection

All posts have been pre-processed, in order to preserve only the elements with an emotional meaning. For this reason, different automatic filters have been used to remove Italian stopwords and punctuation, encode special characters, correct spelling mistakes, substitute contractions with their textual extension, and substitute smiles and emoticons with appropriate words. At the end of this process, sentences have been also filtered using a stemming algorithm, in order to reduce inflected words to their word stem.
The last step consists in the generation of features based on the bag-of-words model, turning each string into a set of attributes representing word occurrences. The term frequency–inverse document frequency (TF-IDF) function has been used to evaluate the relevance of each word, estimating its significance not only in a particular instance, but also in the whole corpus. Finally, the attributes have been filtered using the Information Gain algorithm, in order to extract the most meaningful features.

4.3.4 Classification

In this subsection the approach used to build the hierarchical classifier for the emotion analysis (Figure 4.10) is presented. The manually annotated training set described in Section 4.3.2 has been used to create four different training sets, in which each sentence is labeled according to the task of each classifier:

- **Objectivity / subjectivity classifier**: The training set used is simply the pre-processed training set in which each post associated with an emotion has been labeled as subjective while the remaining ones, for example information requests or communications from the patients' association, have been labeled as objective.
- **Polarity classifier**: The training set used is the pre-processed training set in which all the posts previously labeled as subjective have been divided into positive and negative posts.
- **Positive classifier**: The training set used is composed of the posts previously labeled as positive, divided into love, joy and surprise posts.
- **Negative classifier**: The training set used is composed of the posts previously labeled as negative, divided into anger, fear and sadness posts.

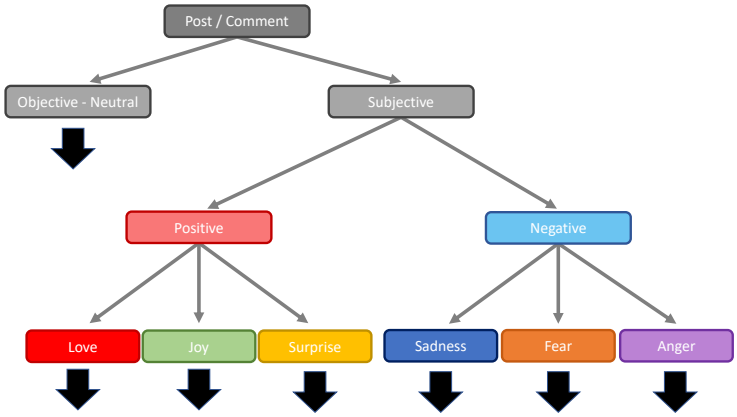


Figure 4.10: Structure of the hierarchical classifier.

4.3.5 Parameter optimization

A preliminary analysis has been performed for each classifier of the seven-output hierarchical classification system. The classifiers were trained using the Naive Bayes

Multinomial algorithm, since it produces the best results. Moreover, some parameters, considered relevant for the training phase, have been optimized. In particular, it has been searched at the same time for (*i*) the optimal length of N-grams to be used as features, and (*ii*) the number of features to select through the Information Gain algorithm. A grid-search optimization, which is simply an exhaustive search through a manually specified subset of the parameters hyperspace of a learning algorithm, has been used in order to select a grid of configurations using cross-validation to estimate the quality of classifiers configured according to them. Results are shown in Table 4.6.

Classifier	N-Gram (max)	Features	Accuracy
Sub/Obj	3	100	91.6
Pos/Neg	2	250	85.6
Fear/Sadness/Anger	3	340	62.3
Love/Joy/Surprise	2	100	70.5

Table 4.6: Parameter optimization results.

4.3.6 Evaluation of the classifier

Table 4.7 reports the results obtained on the test set using the described hierarchical classifier. The total accuracy on the seven emotion classes is 0.489 and in particular the best results in terms of F-Measure have been obtained in the classification of objective and anger classes. Overall negative classes present higher F-Measure values than the positive ones. This is justified, as mentioned previously, with the difficulties in the manual annotation of positive feelings, and, in particular, for the results concerning surprise feeling, due to the lack of available training examples in the group. However, considering the almost complete absence of surprise in the group, these results have not represented a limit to the other following analysis of the dynamics in the group.

4.3.7 Building the Interaction Network

The established relationships among the users in the group discussions have been modeled using a directed and weighted graph, in order to analyze members' activities

Class	F-Measure
Objective	0.69
Sadness	0.43
Fear	0.5
Anger	0.62
Surprise	0.1
Love	0.34
Joy	0.42
Total accuracy : 0.49	

Table 4.7: Evaluation of the hierarchical classifier.

inside the group. A discussion is composed of a Facebook post and its related comments. In the light of this, relationships have been considered unilateral. Considering the possibility introduced by Facebook to leave a comment not only to the post, but also to another comment (a sub-comment), a comment X of a User B to a post or to a comment written by a User A has been considered an interaction of User B directed to User A. A sub-comment is considered an interaction directed only to the author of the comment and not to the author of the post (Figure 4.11). Finally, a global interaction network has been obtained joining all of the networks obtained from the 10,000 discussions in the group.
Analyzing each member's posts and comments through the use of the previously described classifier, each user has been labeled with the most frequent emotion he/she expressed, during the analyzed period. This emotional information has been joined with the interaction network in order to obtain a more detailed graph, where each node represents a member of the group and its color represents its prevailing emotion in the considered period.

4.3.8 Results

In this section the results obtained using the multi-faceted methodology approach are presented. On one hand, these results may provide valuable hints for a better

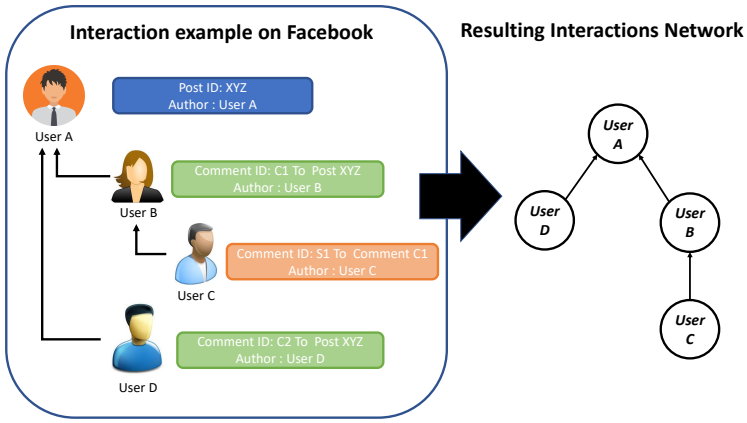


Figure 4.11: Model of interactions.

understanding of the Hidradenitis Suppurativa disease and its impact on patients' lives. On another hand, they also provide useful indications about the importance of a multi-faceted analysis on social media, for avoiding the risk of inferring simplistic results and losing the richness of the underlying complex patterns of social behaviors.

4.3.8.1 Emotion Analysis of Social Media Content

According to the negative impact of this disease [98], the majority of the contents conveys some negative feelings (60%), with sadness being the prevailing emotion (38%). In view of this situation, with the aim of gathering information about the patients' behaviors, an emotion analysis from different perspectives has been carried out on the basis of the following considerations:

- Posts and comments could describe different behaviors and dynamics, thus they have to be analyzed separately. For example, in [124], some of the features used to discern users' behaviors are related to the frequencies of posts and to the fact that a post is being commented.
- A specific analysis of posts and comments, as distinguished and related elements, can give also useful information about supporting, empathy and indif-

ference in the group, depending on the different emotions expressed in a post and its comments.

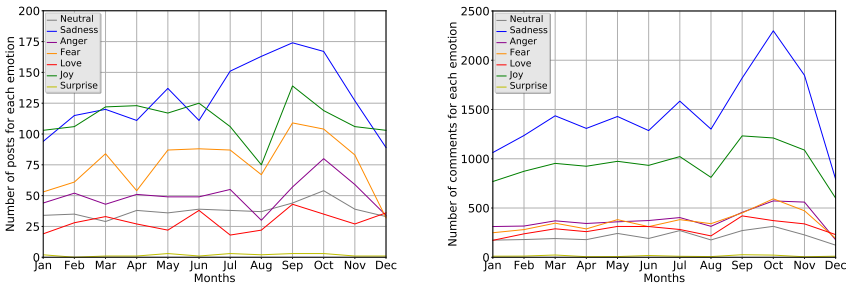
- The publication time of a post could be important. For example, in [64] and [124], micro-blogs expressing sadness are considered as a possible indication of depression, in particular when the publication time is between midnight and 6:00 am.

- Finding possible critical periods of the year, for the management of the patient's disease, requires to search for possible patterns in the monthly distribution of emotions.

The first presented results consist in a comparison between the monthly distributions of the emotions expressed in posts and comments, respectively (Figure 4.12). As reported by Inversa Onlus and as expected in a patients' community, members interact often in the Facebook group when they are in an acute phase of the disease, to vent their frustration and look for emotional support, and less when they feel better. This consideration is important in order to investigate the relationships between the frequency of use of the group and the temporal (monthly) patterns of emotions. It is possible to observe that the month with the highest number of posts and comments is October, while the one with the lowest number is December.
Starting from the previous observation, it is interesting to analyze the differences between emotion trends in comments and posts. In the comments distribution, which can be assumed as a representation of members' reactions in the group, the various emotions show similar trends, during a year. In contrast, considering the same emotions in the post distribution, it is possible to note that trends are much more varied. In particular, during summer, there is a marked increase in the number of posts expressing sadness and a general decrease in posts expressing other feelings. There is also a lower but considerable increase in fear posts during the period preceding summer. In the less active periods of the group, for example in December and January, it is possible to observe that joy is the prevalent emotion among the posts, and there is almost an equality with sadness in the comments distribution. In both cases, the second prevalent feeling during the year is joy and this juxtaposition seems to be related to

the presence inside the group of some influencers who express positive emotions in reaction to other negative posts. These trends are observable and mostly confirmed also in (Figure 4.13), where the same data have been analyzed by using the average percentage of each emotion in the same period and calculating the variance of these values.

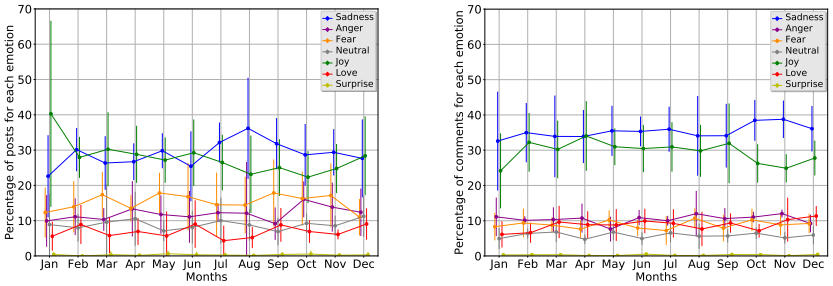
The emotion trends can be read in different ways, but from these results it is conceivable that the most critical period for patients is summer, in which there is not the previously mentioned juxtaposition of emotions among the posts. Thus in summer there is an important prevalence of sad posts, conducting to mostly sad discussions. In fact, comments to a sad post express also sadness in 42% of cases (Table 4.9). An additional hypothesis to justify the absence of positive posts could be that during these months, the most active part in the group remain constituted mostly by sad patients, who probably are mainly suffering some known effects of the disease, caused by higher temperatures and seasonal weather in general.



(a) **(b)**

Figure 4.12: Monthly distributions of emotions expressed in posts (a) and comments (b) over the period 2010-2017.

In the following the results obtained by analyzing posts and comments on the basis of their publication time (Figure 4.14) are presented. Daily hours have been divided in 8 time intervals: Night (00:00 to 03:00), Deep Night (03:00 to 06:00), Early Morning (06:00 to 09:00), Morning (09:00 to 12:00), Lunch time (12:00 to 15:00), Afternoon (15:00 to 18:00), Late Afternoon (18:00-21:00) and Evening (21:00 to 24:00). Like previously described, dynamics between posts and comments are studied separately.



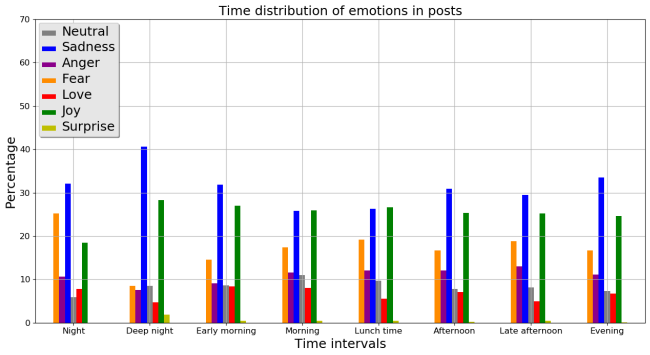
(a) **(b)**

Figure 4.13: The average percentage of emotions expressed in posts (a) and comments (b) for each month over the period 2010-2017 and the related standard deviation.

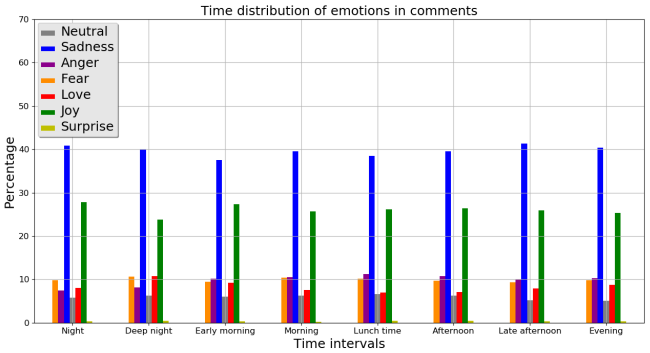
Emotions expressed in the comments seem to be independent from the time, in fact the ratio among emotions is quite constant during the day. Also in this case, the distribution of posts contains the most interesting relationships. In particular, looking at the night intervals, it is possible to observe a considerable increase of fear, that is larger than the joy component. Moreover, although sadness prevails in each interval over other emotions, its peak value is at Deep Night.

The nocturnal relative rise of the negative sentiments (except for anger) could be associated with different factors. However, considering [64] and according to [98], this phenomenon could be related to forms of insomnia and depression.

Analyzing the anonymous data of each member, a matrix of transitions among emotions in the period between 2009 to 2017 has been calculated (Table 4.8). This matrix takes in consideration the transitions of emotions expressed by members year by year. An emotion transition between two years is considered valid for a user, only if the count of all published elements (either posts or comments) during each year is greater or equal to 3. Observing the matrix, it is possible to note that the transitions from fear to joy are more frequent than the transitions from sadness and anger to joy. As a possible interpretation, this result could mean that the worried members are more easily influenced by the rest of the social network. Instead, sadness appears to be the most static emotion. This essentially means that members who express sadness continue to express sadness, regardless of their participation in the group.



(a)



(b)

Figure 4.14: Hourly analysis of emotions in posts and comments, respectively.

4.3.8.2 Social Network Analysis

The second class of results presented in this section concern the analysis of users' relationships and influence factors. These results have been obtained comparing the different dynamics discovered in the Facebook friendship network and in the interaction network, as discussed in Section 4.3. After retrieving information about interactions on Facebook among the group members and classifying posts written by each patient in the group for each year, it is possible to analyze the interaction network in the group and its evolution over the years (Figure 4.15). In particular, at the beginning of the group activities, in 2010, there were few members, that interact a

Table 4.8: Matrix of transitions among emotions. Each row represents a prevailing emotion in a given year, and columns represent the prevailing emotions in the following year, with their probability.

	Joy	**Love**	**Anger**	**Fear**	**Sadness**
Joy	46%	5%	0%	7%	42%
Love	38%	12%	0%	6%	44%
Anger	7%	0%	29%	14%	50%
Fear	21%	0%	0%	3%	76%
Sadness	15%	3%	3%	5%	74%

Table 4.9: Average incidence of emotions expressed in the comments (columns), for each emotion of the commented post (rows).

	Anger	**Fear**	**Sadness**	**Neutral**	**Joy**	**Love**	**Surprise**
Joy	8%	7%	35%	4.2%	34.2%	11%	0.6%
Love	5.4%	5.2%	31.6%	3.3%	31.7%	22.6%	0.2%
Surprise	5.8%	6%	45.6%	5.8%	25%	6%	5.8%
Neutral	9%	9.5%	38.8%	10%	25.7%	6.7%	0.3%
Sadness	9%	9%	42%	5%	26.6%	8%	0.4%
Fear	11%	11.5%	41.7%	6%	23.2%	6.3%	0.3%
Anger	13.7%	10%	37.5%	4%	27.4%	7%	0.4%

lot with each other and express mainly joy. During the following years, the number of members in the group has constantly increased. At the same time, differences about the emotions expressed by isolated or weak-connected members have grown.

Looking at the evolution of the network (Figure 4.15), the green central node with the highest degree, which describes always the same user over the years, remains in evidence. It is clear that, apart from this node, the rest of the connected components express mainly sadness. From this result it is possible to suppose that there is a correlation between the number of connections (node degree) and sadness. In a previous experience [85], analyzing the influence of Facebook friendships until 2016,

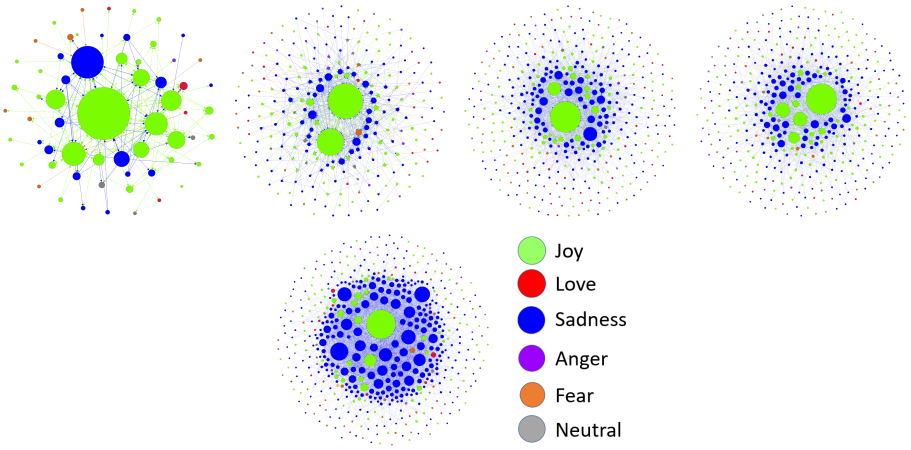


Figure 4.15: Interaction networks in 2010, 2013, 2015, 2016, 2017.

using only the established social graph of the group, friendships resulted to be an important positive emotional influence factor. In light of this, the negative dynamics emerging from the interaction network have been further analyzed, performing a comparison of the two cumulative networks in the period between 2009 and 2017 (Figure 4.16).

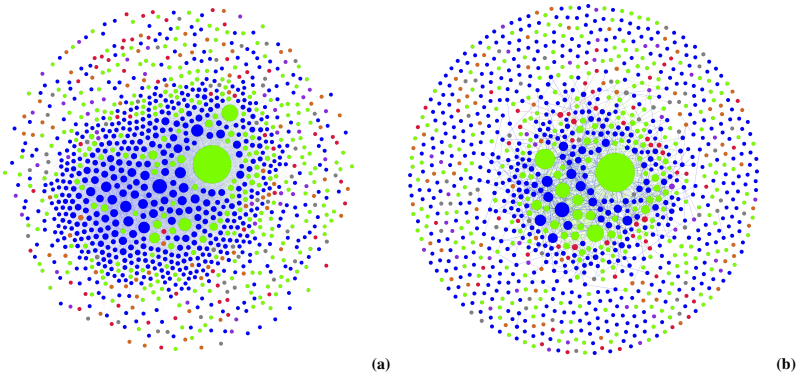


Figure 4.16: Comparison between the friendship network (right) and the interaction network (left) in the period 2009-2017.

In both networks, the size of a node represents its degree. Looking at the distribution of the degree in the two networks (Figure 4.17), it is possible to note that both distributions follow a Power Law function very closely, with an R-Squared value of 0.9845 for the friendship network and 0.9834 for the other one.
It is very interesting to observe that: (*i*) users who interact a lot with each other by writing posts and comments in the group express mainly sadness, while (*ii*) users who establish relationships directly in the form of Facebook friendships express joy. This consideration has been further analyzed by calculating the correlation between friendships, interactions and expressed emotions (Figure 4.18). Another note of interest is that the friendship and the interaction networks are largely different. In fact, only 23.43% of the friendship relations among the group members take place between users who are also connected in the interaction network.

Results presented in this subsubsection have been obtained analyzing how the emotions expressed by patients are correlated with their own degree in the friendship network (i.e., the number of friends in the group) and in the interaction network (i.e., the number of comments written as a response to posts) (Figure 4.18). Values representing interactions and friendships are in two different ranges. For this reason, for the friendship network, members are classified in five categories, according to the number of friendship relations in the group:

- **Zero relationships**
- **Weak**: from 1 to 5 relationships
- **Medium**: from 6 to 15 relationships
- **Moderate**: from 16 to 25 relationships
- **Strong** : 25 or more relationships

Conversely, as regards the interactions, network members are classified in five categories, according to the interactions established by participating to the group discussions in the form of comments (edges' weight in the interaction network):

- **Zero interactions**

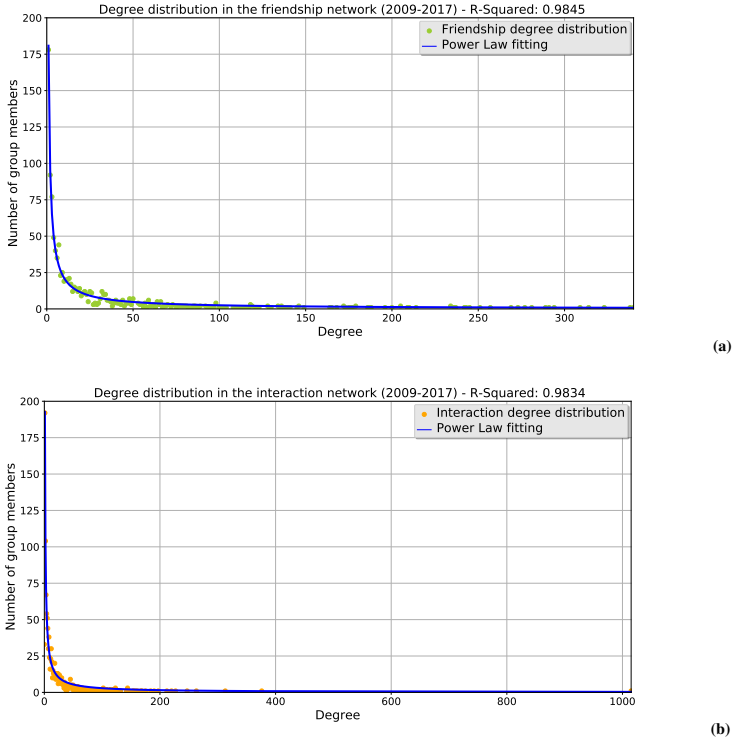


Figure 4.17: Degree distribution in the friendship network and in the interaction network.

- **Weak**: from 1 to 10 interactions
- **Medium**: from 11 to 50 interactions
- **Moderate**: from 51 to 100 interactions
- **Strong** : 101 or more interactions

In the network of friendships, it is worth noting that, for the nodes with higher degree, the predominant emotion is joy, while all the negative emotions decrease. As a possible interpretation, this result may represent an evidence of the positive influence of establishing lasting relationships in the group, also extending outside of it. The

upper graphic in 4.18, for example, shows that all the patients that have developed many relationships with other peers in the group express positive emotions.

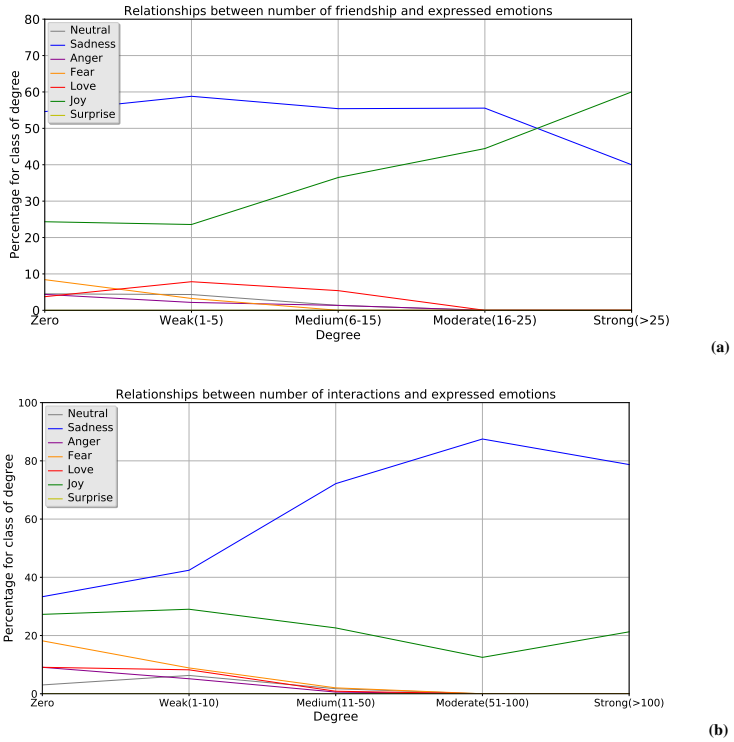


Figure 4.18: Relationships between friendships, interactions and feelings expressed.

Conversely, in the network of interactions, trends describing the two prevalent emotions, sadness and joy, diverge, with an important increase of sadness among the most active nodes in the network.

4.4 Troll Detection in Twitter

Various techniques based on artificial intelligence have been proposed for the automatic detection of online anti-social behaviors, both in existing systems and in the

scientific literature. In this section, TrollPacifier, a holistic system for troll detection, which analyses many different features of trolls and legitimate users on the popular Twitter platform is described. In this system, the most known and promising approaches and research lines are applied, along with original new ideas, in a form that fits such a large public platform. In particular, six groups of features, based respectively on the analysis of writing style, sentiment, behaviors, social interactions, linked media, and publication time have been identified. This study provides: (*i*) the systematic collection and grouping of features, on Twitter; (*ii*) the description of a working holistic system for troll detection, with a very high accuracy (95.5%); and (*iii*) a comparison among the different features, with a machine learning approach. The results demonstrate that automatic classification can be useful in the whole process of identification and management of online anti-social behaviors. However, a multi-faceted approach is required, in order to obtain an adequate accuracy.

4.4.1 TrollPacifier

On the basis of the works analyzed in section 2.2, it can be said that a user has to be considered a troll if his/her activities are driven by an anti-social behavior. Therefore, an ideal approach for identifying such a kind of users makes use of data at the user level. But the considered evaluation also tries to extrapolate additional features from other approaches. Nevertheless, no studies in the literature have tried to comprehensively explore this road. So, another goal of this subsection is to assess the compatibility of the various methods, integrating user-level metrics with features derived from the analysis of published texts and local social graphs.

4.4.2 Actor-based System

In order to implement the TrollPacifier system, ActoDES, which is a software framework which adopts the actor model for simplifying the development of complex distributed systems [14], has been used. Actors are autonomous and concurrent objects, each one characterized by a state and a behavior, and the ability to interact with other agents through the exchange of asynchronous messages. After the analysis of its

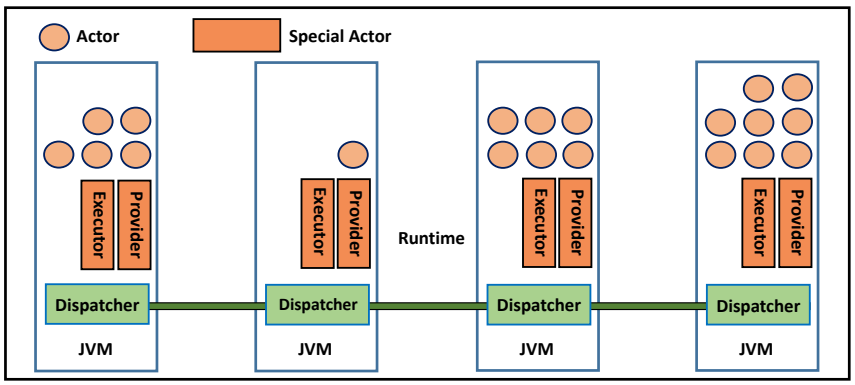


Figure 4.19: Distributed ActoDeS application architecture.

incoming messages, an actor can send more messages to itself or to others, create new actors, update its state, change its behaviors, terminate its own execution, etc. Each behavior can define a policy for handling incoming messages, through handlers called "cases". Each case can only process messages corresponding to a specific pattern. An actor space is intended as a container offering the services needed for the correct execution of a set of actors. In particular, it includes two types of actors: a scheduler and a service provider. The former has the duty to handle the concurrent execution of actors, while the latter provides runtime services, needed by actors to complete their tasks. A subscription service is also available, to facilitate the development of collaborative applications with actors, as shown in [52]. This service has the task to receive incoming messages into a specific mailbox, and forward them to subscriber actors. The actors can eventually handle the messages differently, according to their own behaviors.

Other services, developed for this study, provide additional functionalities to actors, for the continuous analysis of various social streams. In particular, a Twitter service allows other actors to send various kinds of requests:

- User timeline, to obtain the recent tweets of a specified user and save them in a local storage system.

- Content query, to similarly obtain recent tweets published on Twitter and selected according to some constraints specified by the actor.

- Stream, to continuously receive messages published on the platform during the execution; tweets are obtained in the form of JSON objects, which are then stored in a NoSQL repository (namely a MongoDB database [28]).

Leveraging ActoDES and the additional mentioned services, a software system that can be used to track and study a news feed from social media has been built, together with an architecture that can be extended to different cases and also to more complex problems.

4.4.3 Data acquisition

The creation of a dataset of troll users is a crucial point of the analysis.
In order to collect the training set two cascaded approaches have been used. The first one is based on distant supervision [56, 21] and allows one to obtain a raw dataset. The second one consists in manually filtering the previous dataset in order to obtain a more accurate training set.

In the first approach, an idea described by Mihaylov et al. [89] that defines a troll as a user that is called in this way at least N times by N different users has been adapted. Twitter provides a series of official accounts, to which members can report their problems (e.g., @Twitter, @Support, @Safety, @TwitterUK, @TwitterAU, @TwitterSA, etc.). In particular, whenever a common user feels annoyed or even threatened by another, he/she can report the incident to one of these accounts, via a tweet containing a mention to the harasser, in the hope that Twitter administrators take the necessary countermeasures. However, moderators are not always able to take appropriate countermeasures and often many users continue their online activities without being removed.

Therefore, Twitter “Advanced Search” function has been used to select the users who have been reported by other users that accuse them to be trolls (in messages containing words such as “troll”, “ban”, “harass”, “block”, “stalk”, or some common

derivatives). In this way a raw dataset of trolls composed by users mentioned in these messages, but not yet banned by administrators, has been built.

For the non-troll class, the same approach has been used and users starting from general tweets containing common words such as “a”, “an”, “the”, “and” have been selected.

The final obtained dataset is composed of 3000 troll users and 3000 non trolls. By manual inspection of a hundred of instances of this training set, many errors have been found. In fact, a good estimation is that more than a quarter of users mentioned to the support channels do not behave in an anti-social manner.

Therefore, it has been decided to manually filter this raw dataset in order to obtain a more accurate training set, composed of 500 troll and 500 non-troll users, respectively. This final dataset has been validated by multiple independent human judges, through the manual inspection of users reported to the official support channels. In particular, users have been selected after inspecting both their recent timelines and their role and attitude in prolonged discussions, where they were repeatedly mentioned as trolls.

4.4.4 Groups of features

In order to build TrollPacifier 6 groups of features have been identified. They are listed in the following:

- **Sentiment analysis (*SENT*).** This group includes **26 features**, to distinguish positive, negative and objective posts, but also to associate them with more precise emotions. About “sentic computing” [22], TrollPacifier includes the main results of the SenticNet library [24]: sensitivity; polarity; trollness; attention; pleasantness; attitude. Moreover, it takes into account the results of lexicon and rule-based sentiment analysis, using the VaderSentiment library [55]. In particular, from this analysis TrollPacifier gathers values representing the maximum, minimum and average levels of positive, negative and neutral sentiments, polarity and trollness. Additionally, TrollPacifier includes a whole hierarchical emotion detection system, as described in [6]. In particular, the output of each level of classification is used to obtain a feature for the user-level analysis. Thus,

collected features are: number of objective and subjective tweets; number of positive and negative tweets; number of tweets expressing one of the six basic emotions of Parrot's model [102]. Finally, TrollPacifier includes an ad-hoc text-based classifier for evaluating the overall abusiveness of a text, i.e., provoking others to finally report the author as a troll. The classifier is trained with two classes of texts: those written by alleged trolls and those written by normal users.

- **Time and frequency of actions (*TIME*).** This group includes **57 features**, to identify the most active day hours and the time dedicated to each post. Considering the results presented in [27, 36, 90], after an optimization process, features for representing the activity in daily intervals of 4 hours have been included in TrollPacifier. This time interval has been chosen after a thorough comparison, in which automatic classifiers have been trained based on different algorithms (K-nearest neighbors, Naive Bayes, Sequential Minimum Optimization, C4.5) [88] and using different interval durations. Features measuring the activities in intervals of four hours provided consistently the best classification results. In TrollPacifier, the time intervals are distinguished by single day (from Sunday to Saturday), and also grouped together for generic workdays and weekends. In addition to these metrics, additional features consider the frequency of actions in the recent timeline and during the whole user's presence on the platform.

- **Text content and style (*TEXT*).** This group includes **31 features**, to measure the grammatical correctness and the kind of language used in posts. TrollPacifier includes some features for taking into account the readability grades, based on various metrics [33, 27, 89, 37]: Kincaid, ARI, Colemaniau, FleschReadingEase, GunningFogIndex, LIX, SMOGIndex, and RIX. TrollPacifier also includes the following other features in this class: average word length and sentence length, by number of characters, syllables, or words; number of long words and complex words; number of verbs in general and some auxiliary verbs in particular; number of conjunctions, pronouns, prepositions, articles,

subordinations, either in the middle or at the beginning of a phrase; number of hapexes and rare words.

- **User behaviors (*BEHA*).** This group includes **38 features**, to distinguish users participating more actively, i.e., contributing with original messages and media objects. Taking into consideration the experiences of relevant research works [84, 27, 107, 31], a number of features to characterize a user's online behavior have been introduced in TrollPacifier, including: total number of tweets, retweets, replies, favorites, citations and quotes in the timeline; proportion between active actions (original tweets and replies) and passive actions (retweets and quotes); count of various actions associated with a single item (e.g., number of replies to a single tweet); maximum repetitions of a single action.

- **Interactions with the community (*COMM*).** This group includes **34 features**, to highlight a user's integration within his group of followers and followees. To represent a user's relationships within his own community, TrollPacifier includes the following features [80, 107, 79, 99]: number of followers and followees; ratio of these numbers; ratio of tweets per follower; number of posts retweeted or favorited by other users; counts of given and received mentions; number of different mentioned users; counts of different actions, including retweets, replies, mentions, related to a single user or to a single tweet; h-index based on retweets, likes, and their sum.

- **Advertisement of external content (*ADVE*).** This group includes **38 features**, to count the number of references to diverse external content and other channels of discussion. To evaluate the possible usefulness of external links and other forms of advertisement for troll detection [9, 67, 30], the following features have been added to TrollPacifier: number and frequency of URLs in posts and comments, as well as in the profile information provided to the platform; number and frequency of published or advertised videos, images and other media; number and frequency of hashtags.

4.4.5 General feature extraction

Apart from the system-level ActoDES actors, TrollPacifier includes additional actors, as shown in Figure 4.20. They are dedicated to (*i*) basic tasks, like acquiring streaming data and users' profile information from Twitter; (*ii*) direct feature extraction tasks, with different actors for the six different groups of features described in subsection 4.4.4; (*iii*) specialized classification tasks, aimed at calculating additional features through intermediate steps; and (*iv*) final automatic classification, based on different machine learning algorithms. Features are extracted by these actors in both the initialization stage, for creating the training set, and the online operation stage, for evaluating streaming content. Three final classification algorithms have been included: Naive Bayes (NB), Sequential Minimal Optimization (SMO) and Random Forest (RF). The system can also be easily configured to encapsulate any other classification algorithm. As an additional feature, it is also possible to create an online learning loop, thus periodically feeding the training set with newly automatically classified instances, above a certain threshold of confidence [45].

In particular, for the SENT group, some features are obtained through some automatic classifiers that are implemented by few specialized actors integrated into TrollPacifier. One subsystem is dedicated to emotion detection and is built as a hierarchy of classifiers. Another subsystem is dedicated to evaluating the "abusiveness" of a text, through an ad-hoc trained classifier. The role and structure of both subsystems are described in the two following subsections.

4.4.6 Subsystem for emotion evaluation

A subsystem of TrollPacifier is dedicated to the evaluation of the main emotion expressed in a tweet. This classifier is effectively organized in a three-level hierarchy of four specialized classifiers, which reflect a priori relationships between the target emotions. In fact, a common approach to sentiment analysis includes two main classification stages, as defined in chapter 3: definition of objectivity/subjectivity and of polarity of subjectivity.

Extending this basic model, the considered subsystem adds two classifiers as an

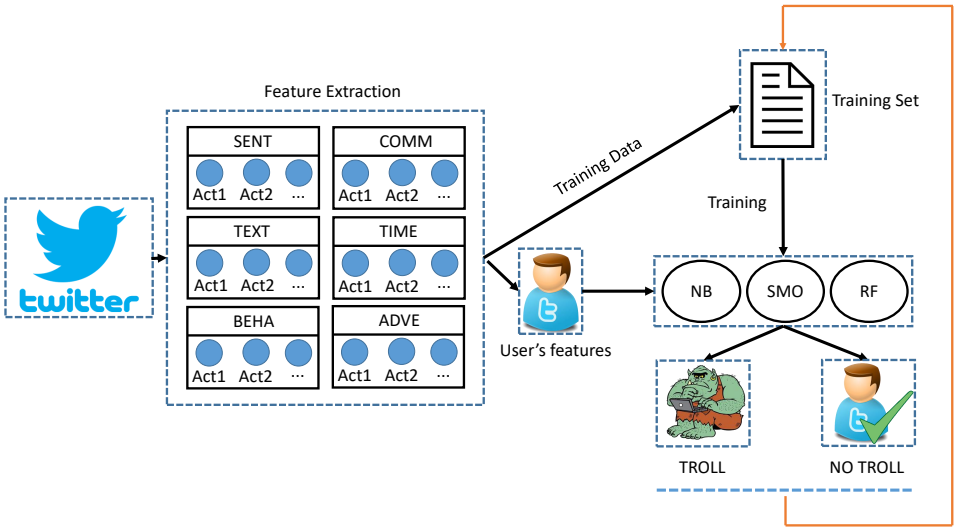


Figure 4.20: Representation of the actor-based system architecture.

additional level for specifying the emotions which characterize subjective tweets, based on Parrott's socio-psychological model [102]. These emotions are analyzed separately by two distinct ternary classifiers, as shown in Figure 4.21.

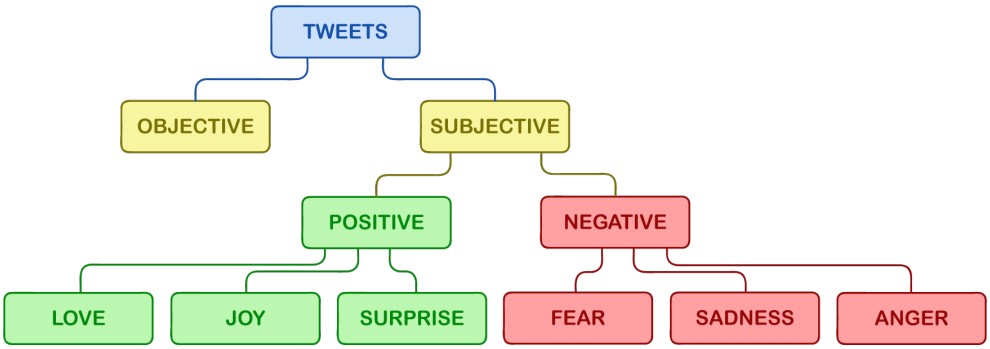


Figure 4.21: Hierarchical emotion classification.

As proven in chapter 3, the a priori domain knowledge embedded into this kind of hierarchical classifier makes it significantly more accurate than a corresponding 7-output flat classifier.

4.4.7 Subsystem for abusiveness evaluation

With the term abusiveness one can consider an online behavior characterized by improper or wrongful use, provoking others to finally report the author as a troll. The classifier is trained with two classes of texts: those written by alleged trolls and those written by normal users. In particular, the messages used for training the ad-hoc classifier are exactly all the posts of the 6000 users described in subsection 4.4.3. As a first step, the collected 542,676 posts have been cleaned following the techniques used in [7] (text conversion to lowercase, white space stripping, Stemming, English stop words removal etc.), in order to increase the final accuracy. The training set is balanced (same number of troll posts and non-troll posts).

Since the classification of text documents requires a proper text representation, in this research the bag-of-words model, which is simply the extraction of all words of the corpus and the representation of each sentence as the vector of the corresponding occurrences, has been chosen.

Due to the large number of features of the bag-of-words model (corresponding to the number of different words used in the corpus), it is necessary to reduce the model complexity, retaining only the most discriminant features. Features have been selected according to the Information Gain (IG) criterion [88], to improve accuracy and reduce the time required by the learning step. Information Gain computes the expected entropy reduction by measuring the amount of *a priori* information about the class prediction when the only information available is the presence of a feature and its corresponding class distribution.

In this study, Information Gain has been computed for each feature. Then, the IG scores of all the attributes have been ranked in descending order; finally, in the learning step, only the top k have been considered. It is to be noticed that the k value has been optimized with a grid search optimization method [16].

Grid search is simply an exhaustive searching through a manually specified subset

of the hyperparameter space of a learning algorithm and it is usually guided by a performance metric (in this case the classification accuracy). After the grid search optimization process, the best value found for k is about 30,000 (number of features).

It is to be noticed that the previously described methods (bag-of-words model, information gain, grid search optimization, etc.) are standard approaches for the creation of an automatic text classification system [126].

The created classification model is then used for the evaluation of the "abusiveness" feature, which corresponds to the percentage of troll posts published by a user with respect to the total number of his posts.

4.4.8 Results

The experimental results described in this subsection show the importance of the considered features for the automatic detection of troll users. The results are presented in three separate sections, in order to highlight the effectiveness of the six considered groups of features (COMM, TEXT, BEHA, SENT, TIME, ADVE), the contribution of each feature individually, and the execution time. The results obtained by considering a dataset containing all the features of the six considered groups (TOT dataset) have been also analyzed.

Regarding the classification methods, it has been decided to show the results of the 3 best classification algorithms [88] among those tested: Sequential minimal optimization (SMO), Naive Bayes (NB), and Random Forest (RF).

4.4.8.1 Comparison of groups of features

Table 4.10 shows the different accuracies obtained with 10 runs of 10-folds cross validation on different datasets and different classification algorithms. Ten runs of 10-folds cross validation have been performed in order to obtain more reliable results. The first six datasets (SENT, TIME, TEXT, BEHA, COMM, ADVE) are obtained from the one described in 4.4.3, by selecting only the features of the corresponding group. The TOT dataset is exactly the same described in 4.4.3 and shows the accuracy of the system by considering all the features.

Table 4.10: Accuracy, F-measure, Kappa statistic, AUC (Area Under the Receiver Operating Characteristic curve) and Recall for each dataset, using different Machine Learning algorithms (SMO, NB, RF).

	Accuracy (%)			F-measure			Kappa			AUC			Recall		
	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF	SMO	NB	RF
SENT	78.80	72.31	78.97	0.79	0.71	0.79	0.58	0.45	0.58	0.79	0.79	0.87	0.77	0.67	0.79
TIME	67.84	61.28	75.65	0.57	0.41	0.75	0.36	0.23	0.51	0.68	0.77	0.83	0.44	0.27	0.74
TEXT	67.56	64.97	68.47	0.66	0.63	0.69	0.35	0.30	0.37	0.68	0.69	0.74	0.64	0.59	0.70
BEHA	75.88	58.62	79.59	0.75	0.70	0.80	0.52	0.17	0.59	0.76	0.82	0.87	0.70	0.97	0.79
COMM	80.45	74.96	83.16	0.80	0.72	0.83	0.61	0.50	0.66	0.80	0.83	0.91	0.78	0.64	0.83
ADVE	78.07	71.70	85.01	0.78	0.67	0.85	0.56	0.43	0.70	0.78	0.79	0.92	0.76	0.59	0.85
TOT	95.52	80.25	88.28	0.95	0.78	0.88	0.91	0.60	0.77	0.96	0.90	0.96	0.95	0.69	0.89

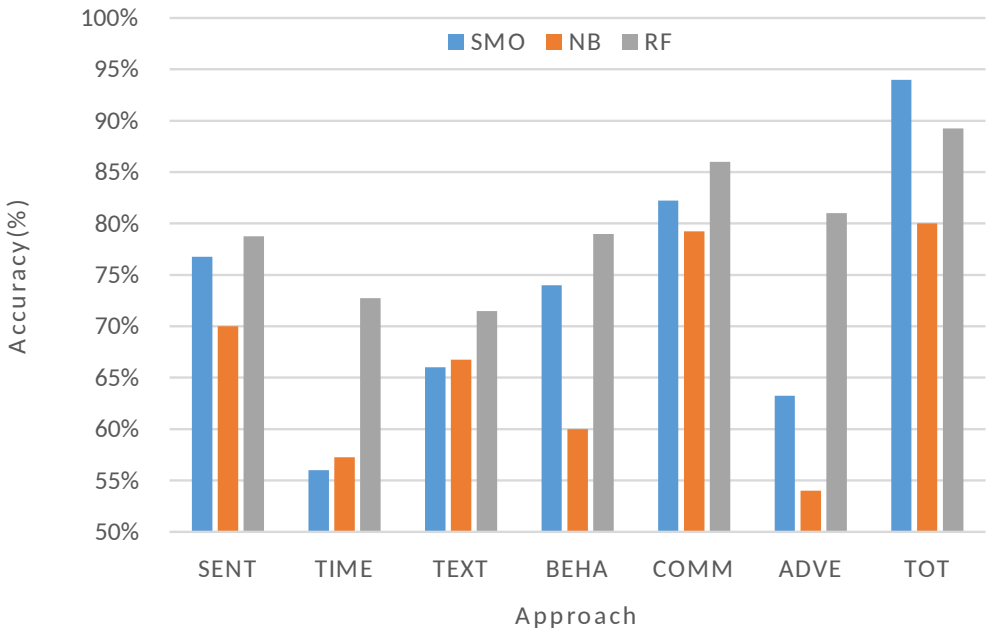


Figure 4.22: Accuracy obtained with different groups of features and different algorithms.

A general aspect that is deduced from the results is that some groups of metrics work better than others to distinguish the considered classes. In particular, the

Table 4.11: Accuracy, F-measure, Kappa statistic, AUC (Area Under the Receiver Operating Characteristic curve) and Recall obtained by removing one group of features at a time.

	Accuracy (%)			F-measure			Kappa			AUC			Recall		
	SMO	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**
TOT-SENT	94.25	79.67	87.38	0.94	0.77	0.87	0.89	0.59	0.75	0.94	0.89	0.95	0.93	0.68	0.88
TOT-TIME	95.07	82.22	89.27	0.95	0.84	0.89	0.90	0.64	0.79	0.95	0.89	0.96	0.94	0.92	0.89
TOT-TEXT	94.53	79.36	88.26	0.94	0.76	0.88	0.89	0.59	0.77	0.95	0.89	0.96	0.93	0.67	0.89
TOT-BEHA	95.05	73.87	88.98	0.95	0.67	0.89	0.90	0.48	0.78	0.95	0.90	0.96	0.94	0.54	0.90
TOT-COMM	90.40	75.51	86.92	0.90	0.71	0.87	0.81	0.51	0.74	0.90	0.88	0.95	0.89	0.59	0.87
TOT-ADVE	89.49	77.29	86.26	0.89	0.74	0.86	0.79	0.55	0.73	0.89	0.89	0.94	0.89	0.63	0.87
TOT	95.52	80.25	88.28	0.95	0.78	0.88	0.91	0.60	0.77	0.96	0.90	0.96	0.96	0.69	0.89

Community and the Advertisement group perform better than the others.

Moreover, Random Forest allows one to achieve the highest accuracy for all groups, but it is outperformed by SMO using the TOT dataset. In fact, in [38] it was demonstrated that the RF algorithm does not have high performance when dealing with high-dimensional data (like the TOT case, which clearly includes much more features than any individual group), especially in presence of dependencies.

It is also interesting to notice that Naive Bayes is often outperformed by the other two classification algorithms. Probably, this is due to the strong dependence among the features inside the same group, which are considered independent by the Naive Bayes assumption. The results can be better appreciated by looking at Figure 4.22.

In order to better highlight the importance of each group, it has been decided to evaluate complementary combinations of features. In particular, in Table 4.11 the first six datasets (TOT-SENT, TOT-TIME, TOT-TEXT, TOT-BEHA, TOT-COMM, TOT-ADVE) are obtained from the TOT dataset by removing the features of the corresponding group. The results are also described in Figure 4.23.

In addition to the evaluations shown in Table 4.11 and Figure 4.23, it has also been tried to combine the contribution of each group in each classification algorithm with an ensemble learning method [114], in an effort to achieve better accuracy. The main premise of ensemble learning is that, by combining multiple models, the errors of a single classifier will be probably compensated by other classifiers, and, as a result, the overall prediction performance of the ensemble would be better than that of a

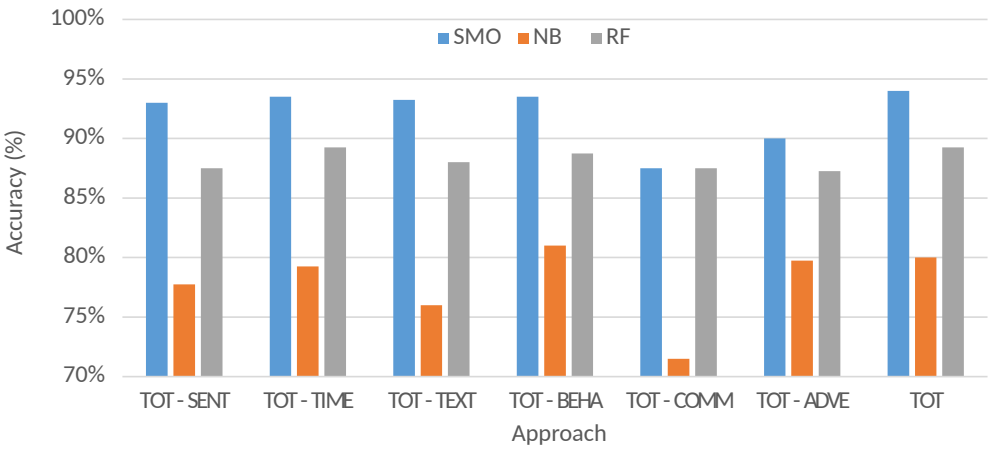


Figure 4.23: Results obtained by removing one group of features at a time.

single classifier. In particular, the prediction confidences of each classifier for each group have been combined using a stacking model [123] with a neural network as a meta learner (Figure 4.24). The hyperparameters of the neural network have been optimized using a grid search optimization method. In the optimal configuration, the network achieves an accuracy of 93,6%, which is lower than the accuracy obtained by the SMO classifier using all the features. This is probably due to the dependencies among features of different groups, which cannot be identified by the network since the inputs are only the confidence levels of previous classification algorithms.

4.4.8.2 Single feature analysis and remarks

At a finer level of analysis, it is possible to assess which features have the largest influence on the results. In particular, the Information Gain algorithm has been used to find the most relevant features. IG evaluates the worth of an attribute by calculating the reduction in entropy for each feature. The result is that features that perfectly discriminate the class give maximal information and unrelated features give low information.

Table 4.12 describes the first 10 features in decreasing order of Information Gain,

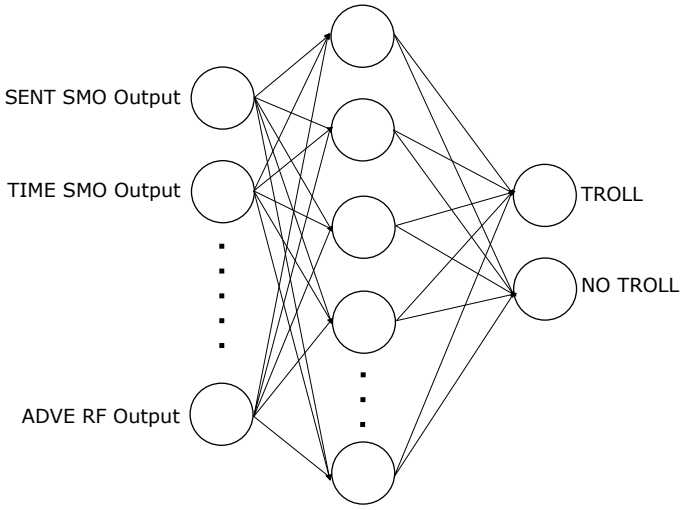


Figure 4.24: The neural network meta learner in the stacking ensemble learning model.

together with the corresponding group.

Features from the COMM group are the most discriminating ones (4 of the best 10 features belong to this group). However, it is to be noticed that features from different groups provide important contributions for automatic classification and also provide useful insights about diverse aspects of online trolling. In particular, these are the most discriminating features for each group:

- COMM. The most valuable contribution, in absolute, is provided by features based on the number of mentions in the timeline. In the same group, other important features are based on the attention given to other accounts, measured on the basis of continued interactions. This indicates that troll users tend to engage in multiple and long conversations, probably due to prolonged arguments with other users. The number of followers is also discriminating, as a troll user is generally not well received by the community. The typical low level of success also leads to fewer tweets that are re-shared or liked by other users.

- SENT. Among the features measuring sentiments and emotions of tweets, abu-

Table 4.12: The first 10 features in decreasing order of Information Gain.

Description	Group	IG
#Users mentioned in the quoted tweets + #Users mentioned in the tweets	COMM	0.327
The results of an ad-hoc text-based classifier for evaluating the "abusiveness" of a text (described in 4.4.4)	SENT	0.275
#Answers to the user's tweets + #Retweets + #Shares	COMM	0.250
#Public lists the user belongs to	ADVE	0.235
#Followers of the user	COMM	0.228
The frequency of user's messages on Mondays from 00:00 to 04:00 (4.4.4)	TIME	0.139
#Urls in posts and comments	ADVE	0.191
#Replies to the user	COMM	0.178
The frequency of user's messages on Thursdays from 08:00 to 12:00 (4.4.4)	TIME	0.167
The frequency of user's messages on Fridays from 08:00 to 12:00 (4.4.4)	TIME	0.164

siveness is the most discriminative. In fact, it is based on an automatic classifier trained with messages written by users reported to the support channels of Twitter. This means that the lexicon used by trolls is quite distinguishing. Other features in this group provide less important contributions. The fact that trolls are not strongly characterized by emotions can be a manifestation of their Machiavellianism, which is associated with the personality of online trolls [19].

- ADVE. Generally, a troll has little incentive to subscribe to lists on Twitter, which are mainly used to remain informed on a specific topic. Instead a troll tend to publish more URLs and to reshare more tweets from various sources, indicating that some trolls may be effectively engaged in various types of campaigns. They also use more hashtags, possibly to gain visibility and because they deal with multiple and diverse topics, thus lacking focus.

- TIME. It is quite interesting that the simple analysis based only on the daily and hourly frequency of messages provide quite good results. In fact, a troll produces many more tweets than a normal account, in particular deep in the night. This can be related to availability of time and to prolonged arguments, but it can also be related to personality traits of online trolls which would deserve further studies [70].

- BEHA. While patterns of behavior are generally useful for bot detection, instead they provide minor gains for troll detection. Some features in this group, based on the number of replies to other tweets and other users, indicate an attitude of trolls to follow and engage in multiple conversations. In fact, triggering conflicts with other users result in verbal crossfires that go longer than a normal conversation.

- TEXT. Among the metrics based on the text of the tweets, the most discriminating are related to the indices of readability. This study confirms that troll users tend to write less readable posts, as they pose less care in the drafting of their texts [27]. Other relevant features in this group include the use of emoticons, the richness of vocabulary and the number of hapaxes, i.e., words appearing only once in a user's tweets.

4.4.8.3 Execution time

Finally, to evaluate the applicability of the proposed system in real contexts, the execution time for both downloading and analyzing data has been measured. In particular, for downloading the tweets to analyze, the average time required, by user, is 1.748 s, with a standard deviation of 0.298 s. Instead, in order to analyze data and then provide a user's actual features, the average time required is 43.819 s, with a standard deviation of 40.921 s. These aggregated results have been obtained from tests executed for many dozens of different users. They refer to the current implementation, which may be certainly improved through optimization and parallelization, running on a desktop PC with an i5-4210U processor, 16 GB of ram, SSD.

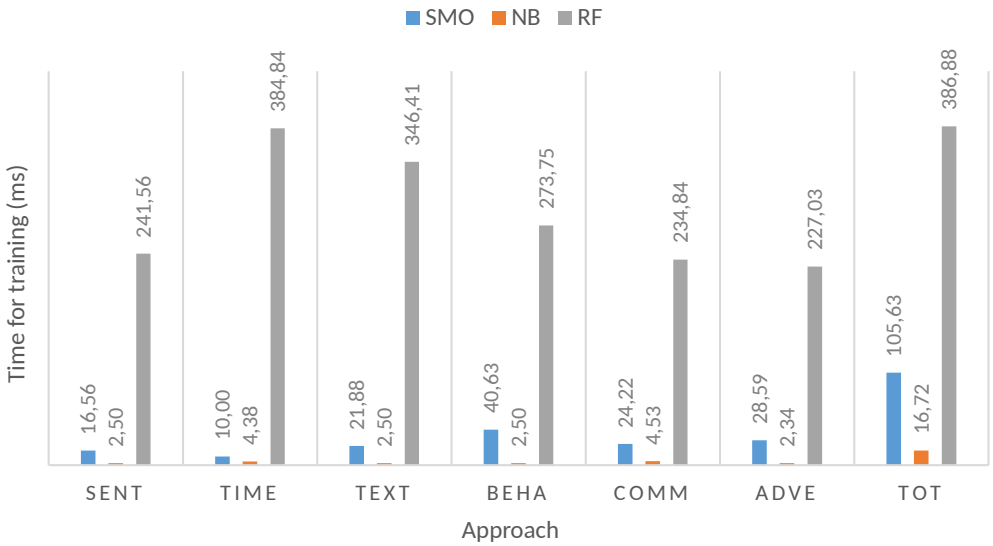


Figure 4.25: Time required to train the classifiers.

After having calculated all features, the time required for actual classification is practically negligible for all evaluated algorithms. In fact, the average value is 3 ms, with a standard deviation of 6 ms. To highlight some differences among the classification algorithms, the time required for training has been also evaluated, with results shown in Figure 4.25. It is worth underlying that the training process happens essentially offline, after acquiring the training set and before starting the system. However, this evaluation can be useful in the realization of more adaptable systems. In fact, in TrollPacifier, it is also possible to collect new training data at runtime to perform online learning, *i.e.*, (*i*) enrich the training set with some new instances observed while running the system, and then (*ii*) periodically update the classification model, by repeating the training process. In this case, the different computational weight of the training phase can also be taken into account.

Chapter 5

Conclusions

This chapter presents the conclusions of the whole results obtained throughout this thesis.

As regards pre-processing of the datasets, this is an essential phase in all relevant applications of data mining. In Sentiment Analysis, in particular, it is cited in virtually all available research works. However, few works have been specifically dedicated to understanding the role of each one of the basic pre-processing techniques, which are often applied to textual data. Therefore, one of the contributions of this research thesis has been to provide a more precise measure of the impact of these basic techniques that can improve the knowledge of the whole data mining process. As an interesting obtained result, it is worth noting that using a dictionary did not enhance the performances in the considered tests, but it increased the elaboration-time needed for cleaning raw data. All other techniques, instead, provided significant improvements to the classifier performances. Some of the techniques simply removed useless noise in the raw data, while others increased the relevance of some concepts, reducing similar terms and expression forms to their most basic meaning. The considered research has been conducted over data which originated from Twitter; however, a similar analytical work could be performed on different kinds of data sets, to have a more comprehensive understanding of the different pre-processing filters. The decision to mix some of these filters is often correct. However, it should be better motivated by empirical data and

result evaluations for various application domains and the peculiar nature of their textual data.

The techniques mentioned above have also been used to obtain a polished dataset for sentiment analysis. As a matter of fact, for the creation of a classifier for emotion detection, it is of utmost importance the collection of a proper training set with low costs and efforts; so, an approach for automatically deriving a training set is essential. Even if it has been proven that training sets obtained with distant supervision correspond well to annotation of human judges, in the this thesis it has been shown that it is possible to increment the quality of the training set using a simple and automated dataset pruning technique.

In the thesis the problem of automatic classification of tweets, according to their emotional value, has also been tackled considering Parrott's model of six primary emotions and the comparison of a flat classifier with a hierarchical classifier. The performed tests has demonstrated that the domain knowledge embedded into the hierarchical classifier makes it more accurate than the flat classifier. Moreover, the obtained results have proven that the process of automatic construction of training sets is viable, at least for sentiment analysis and emotion classification, since the automatic filtering of the training data makes it possible to create training sets that improve the quality of the final classifier with respect to a "blind" collection of raw data based only on the hash-tags. The results that have been obtained are comparable with those found in similar works in the current literature.

As concerns the results obtained from the synthesis of Social Network Analysis and Sentiment Analysis, one of the considered approach has been tested on the #SamSmith channel during the Grammy Awards in 2015, and on the #Ukraine channel during the 2014 crisis. The implemented methodology has allowed to get a training set for the classifiers that deal with Sentiment Analysis, and to make a thorough study of the network topology. The study of the global sentiment within the network has highlighted the typical problems of Sentiment Analysis (irony, sarcasm, lack of information, etc.). Additionally, some peculiar problems of the considered channels have also been detected (such as the quotes of songs). The performances obtained by the classifiers during the tests conducted on the training set and the analysis of the

case studies have shown, however, good and promising results.

Some of the results of the research carried out about support groups in Facebook, though hoped, were not so obvious to the group representatives. In fact, the obtained results may be useful both for the analyzed group and also for the study of online support communities, in general. A suggestive result of the research is the recognition of the presence of a significant correlation between the degree of a node and the prevalence of positive emotions, indicating a possible positive role of building stable social relations inside the support group. Other impressive results have emerged only though the comprehensive analysis we carried out, as a manifestation of the complex nature of online social dynamics. Very different results can be obtained by studying posts and comments, separately, as well as confronting the social network built on friendship relationships with one based on interactions in the group. In this sense, the proposed combined approach provides some interesting insights that should be further analyzed, in different online groups. In future researches, it would be possible to perform a similar analysis in other online groups of patients, of different diseases, or to compare the current findings with results obtained by other future research works. This way, it will be possible to verify if some kinds of social and psychological dynamics are common to similar online communities.

As regards the research about troll detection, it can be said that the identification of troll users is possible. Some of the techniques present in the literature are described as able to obtain significant results, but usually in much smaller and controllable environments than the one chosen in the research carried out in this thesis. In fact, also in a large and dynamic context like Twitter, the applicability of some techniques described in the scientific literature has been successfully verified. However, it is also evident that currently exploited methodologies can be significantly improved since many works rely only on specific aspects of users' online presence. The fusion of different types of metrics is possible and desirable since the problem of troll detection is complicated by its nature, as a strong subjectivity of the act characterizes it.
Considering that the dimensions along which the online trolling phenomenon develops are numerous and various, it has been proven that some methodological and practical guidelines can be followed. In particular, the studied methodology has been applied

to Twitter, as a very popular microblogging platform. The considered metrics and algorithms are especially tailored for this platform. In the future, it is planned to extend this research work to different scenarios, since this research poses good basis for a more comprehensive understanding of the problem and the value of its multifaceted aspects, for building useful automatic classification tools and thus improving the conditions for more participatory online communities.

As regards general future developments, the methodologies presented in the thesis could be also applied, in the future, to novel types and scenarios of social networks, such as those ones based on block-chain, or those ones composed of only smart devices of the Internet of Things. Moreover, the presented sentiment analysis could be integrated with a more thorough social network analysis, in order to detect particular graph motifs in the interaction graph of the users and, as a consequence, to connect them to the sentiments expressed in the posts/comments themselves.

Chapter 6

Papers from this thesis

In this chapter we highlight the publications coming from the work of this thesis. They are the following:

1. Paolo Fornacciari, Monica Mordonini and Michele Tomauiolo, “Social Network and Sentiment Analysis on Twitter: Towards a Combined Approach”, Proceedings of the 1st International Workshop on Knowledge Discovery on the WEB, Cagliari, Italy, September 3-5, 2015.

2. G. Angiani et al., “A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter”, Proceedings of the 2nd International Workshop on Knowledge Discovery on the WEB, Cagliari, Italy, September 8 - 10, 2016.

3. Angiani G., Cagnoni S., Chuzhikova N., Fornacciari P., Mordonini M., Tomaiuolo M. (2016), “Flat and Hierarchical Classifiers for Detecting Emotion in Tweets”. In: Adorni G., Cagnoni S., Gori M., Maratea M. (eds) AI*IA 2016 Advances in Artificial Intelligence. AI*IA 2016. Lecture Notes in Computer Science, vol 10037. Springer, Cham

4. Lombardo G., Ferrari A., Fornacciari P., Mordonini M., Sani L., Tomaiuolo M. (2018), “Dynamics of Emotions and Relations in a Facebook Group of Patients with Hidradenitis Suppurativa”. In: Guidi B., Ricci L., Calafate C., Gaggi

O., Marquez-Barja J. (eds) Smart Objects and Technologies for Social Good. GOODTECHS 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 233. Springer, Cham

5. Sani L., Lombardo G., Pecori R., Fornacciari P., Mordonini M., Cagnoni S. (2018) "Social Relevance Index for Studying Communities in a Facebook Group of Patients". In: Sim K., Kaufmann P. (eds) Applications of Evolutionary Computation. EvoApplications 2018. Lecture Notes in Computer Science, vol 10784. Springer, Cham

6. Cagnoni S. et al. (2018), "Automatic Creation of a Large and Polished Training Set for Sentiment Analysis on Twitter". In: Nicosia G., Pardalos P., Giuffrida G., Umeton R. (eds) Machine Learning, Optimization, and Big Data. MOD 2017. Lecture Notes in Computer Science, vol 10710. Springer, Cham

7. Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, Michele Tomaiuolo, "A holistic system for troll detection on Twitter", Computers in Human Behavior, Volume 89, 2018, Pages 258-268.

8. Lombardo, G., Fornacciari, P., Mordonini, M. et al., "A combined approach for the analysis of support groups on Facebook - the case of patients of hidradenitis suppurativa", Multimedia Tools and Applications (2019). In press.

Bibliography

[1] A. Addis, G. Armano, and E. Vargiu. A progressive filtering approach to hierarchical text categorization. *Communications of SIWN*, 5:28–32, 2008.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passoneau. *Sentiment Analysis of Twitter Data*. Computer Science - Columbia University (New York, USA), 2011.

[3] D. Al-Hajjar and A. Z. Syed. Applying sentiment and emotion analysis on brand tweets for digital marketing. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on*, pages 1–6. IEEE, 2015.

[4] L. Allisio, V. Mussa, C. Bosco, V. Patti, and G. Ruffo. Felicittà: Visualizing and estimating happiness in italian cities from geotagged tweets. In *ESSEM@ AI* IA*, pages 95–106, 2013.

[5] M. Amoretti, A. Ferrari, P. Fornacciari, M. Mordonini, F. Rosi, and M. Tomaiuolo. Local-first algorithms for community detection. In *KDWeb 2016, 2nd International Workshop on Knowledge Discovery on the WEB*, 2016.

[6] G. Angiani, S. Cagnoni, N. Chuzhikova, P. Fornacciari, M. Mordonini, and M. Tomaiuolo. Flat and hierarchical classifiers for detecting emotion in tweets. In *AI* IA 2016 Advances in Artificial Intelligence*, pages 51–64. Springer, 2016.

[7] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, and S. Manicardi. A comparison between preprocessing techniques for sentiment analysis in twitter. In *KDWeb*, pages 1–11, 2016.

[8] G. Armano, F. Mascia, and E. Vargiu. Using taxonomic domain knowledge in text categorization tasks. *International Journal of Intelligent Control and Systems, special issue on Distributed Intelligent Systems*, 2007.

[9] J. Aro. The cyberspace war: propaganda and trolling as warfare tools. *European View*, 15(1):121–132, Jun 2016.

[10] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[11] A. Balahur. *Sentiment Analysis in Social Media Texts*. European Commission Joint Research Center (Varese, Italy), 2013.

[12] M. Baldoni, C. Baroglio, V. Patti, and P. Rena. From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54, 2012.

[13] Y. Bao, C. Quan, L. Wang, and F. Ren. The role of pre-processing in twitter sentiment analysis. In *International Conference on Intelligent Computing*, pages 615–624. Springer, 2014.

[14] F. Bergenti, A. Poggi, and M. Tomaiuolo. An actor based software framework for scalable applications. *Lecture Notes in Computer Science (LNCS)*, 8729:26–35, 2015. Proc. 7th International Conference on Internet and Distributed Computing Systems (IDCS 2014); Calabria; Italy; 2014-09-22/24 [MT].

[15] K. Berger, J. Klier, M. Klier, and F. Probst. A review of information systems research on online social networks. *Communications of the Association for Information Systems*, 35(1):145–172, 2014.

[16] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, Feb. 2012.

[17] O. Bogolyubova, P. Panicheva, R. Tikhonov, V. Ivanov, and Y. Ledovaya. Dark personalities on facebook: Harmful online behaviors and language. *Computers in Human Behavior*, 78:151–159, 2018.

[18] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11:450–453, 2011.

[19] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102, 2014.

[20] S. L. Buglass, J. F. Binder, L. R. Betts, and J. D. Underwood. Looking for trouble: A multilevel analysis of disagreeable contacts in online social networks. *Computers in Human Behavior*, 70:234–243, 2017.

[21] S. Cagnoni, P. Fornacciari, J. Kavaja, M. Mordonini, A. Poggi, A. Solimeo, and M. Tomaiuolo. Automatic creation of a large and polished training set for sentiment analysis on twitter. In *International Workshop on Machine Learning, Optimization, and Big Data*, pages 146–157. Springer, 2017.

[22] E. Cambria, P. Chandra, A. Sharma, and A. Hussain. Do not feel the trolls. In *CEUR Workshop Proceedings*, volume 664, pages 1–12, 2010.

[23] E. Cambria, D. Olsher, and D. Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[24] E. Cambria, S. Poria, D. Hazarika, and K. Kwok. Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*, pages 1795–1802, 2018.

[25] J. Chen, H. Huang, S. Tian, and Y. Qu. Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, 36(3, Part 1):5432 – 5435, 2009.

[26] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1217–1230. ACM, 2017.

[27] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. *arXiv preprint arXiv:1504.00680*, 2015.

[28] K. Chodorow. *MongoDB: the definitive guide*. " O'Reilly Media, Inc.", 2013.

[29] M. Clément and M. J. Guitton. Interacting with bots online: Users' reactions to actions of automated programs in wikipedia. *Computers in Human Behavior*, 50:66–75, 2015.

[30] J. Cordi. *Social Media Revolution: Political and Security Implications*. NATO Parliamentary Assembly, 2017.

[31] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31(5):58–64, 2016.

[32] A. Dance. Communication: Antisocial media. *Nature*, 543(7644):275–277, 2017.

[33] J. de-la Pena-Sordo, I. Santos, I. Pastor-López, and P. G. Bringas. Filtering trolling comments through collective classification. In *International Conference on Network and System Security*, pages 707–713. Springer, 2013.

[34] C. Dellarocas. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593, 2006.

[35] L. Derczynski and K. Bontcheva. Pheme: Veracity in digital social networks. In *UMAP Workshops*, pages 1–4, 2014.

[36] N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 133–142. ACM, 2011.

[37] I. O. Dlala, D. Attiaoui, A. Martin, and B. B. Yaghlane. Trolls identification within an uncertain framework. In *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, pages 1011–1015. IEEE, 2014.

[38] T.-N. Do, P. Lenca, S. Lallich, and N.-K. Pham. Classifying very-high-dimensional data with random forests of oblique decision trees. In *Advances in knowledge discovery and management*, pages 39–55. Springer, 2010.

[39] S. Dollberg. The metadata troll detector. *Tech. Rep. Semester Thesis*, 2015.

[40] J. S. Donath et al. Identity and deception in the virtual community. *Communities in cyberspace*, 1996:29–59, 1999.

[41] P. Ducange, R. Pecori, and P. Mezzina. A glimpse on big data analytics in the framework of marketing strategies. *Soft Computing*, pages 1–18, 2017.

[42] S. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM, 2000.

[43] B. Duncan and Y. Zhang. Neural networks for sentiment analysis on twitter. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on*, pages 275–278. IEEE, 2015.

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

[45] P. Fornacciari, M. Mordonini, A. Poggi, and M. Tomaiuolo. Software actors for continuous social media analysis. *CEUR Workshop Proceedings*, 1867:84–89, 2017.

[46] P. Fornacciari, M. Mordonini, and M. Tomauiolo. Social network and sentiment analysis on twitter: Towards a combined approach. In *KDWeb 2015, 1st International Workshop on Knowledge Discovery on the WEB*, 2015.

[47] E. Franchi, A. Poggi, and M. Tomaiuolo. Blogracy: A peer-to-peer social network. *International Journal of Distributed Systems and Technologies (IJDST)*, 7(2):37–56, 2016.

[48] E. Franchi, A. Poggi, and M. Tomaiuolo. Social media for online collaboration in firms and organizations. *International Journal of Information System Modeling and Design (IJISMD)*, 7(1):18–31, 2016.

[49] E. Franchi and M. Tomaiuolo. Distributed social platforms for confidentiality and resilience. *Social Network Engineering for Secure Web Data and Services*, page 114, 2013.

[50] V. Francisco, P. Gervás, and F. Peinado. Ontological reasoning to configure emotional voice synthesis. In *Procs of Web Reasoning and Rule System*. Springer, 2007.

[51] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[52] S. Gallardo-Vera and E. Nava-Lara. Developing collaborative applications with actors. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1, pages 1–5, 2015.

[53] H. Gao, J. Zhu, and C. Li. The analysis of uncertainty of network security risk assessment using dempster-shafer theory. In *Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on*, pages 754–759. IEEE, 2008.

[54] D. Ghazi, D. Inkpen, and S. Szpakowicz. Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 140–146. Association for Computational Linguistics, 2010.

[55] C. H. E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225, 2014.

[56] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[57] M. Golf-Papez and E. Veer. Don't feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, 33(15-16):1336–1354, 2017.

[58] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf*, 22(3):251–255, 2013.

[59] J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank. Online social networking by patients with diabetes: a qualitative evaluation of communication with facebook. *Journal of general internal medicine*, 26(3):287–292, 2011.

[60] E. Haddi, X. Liu, and Y. Shi. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32, 2013.

[61] C. Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions, 2010.

[62] C. Healey and S. Ramaswamy. Visualizing twitter sentiment, 2010. Online; Accessed on 6-17-2016.

[63] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab. Searching for safety online: Managing" trolling" in a feminist forum. *The Information Society*, 18(5):371–384, 2002.

[64] X. Huang, L. Zhang, T. Liu, D. Chiu, T. Zhu, and X. Li. Detecting suicidal ideation in chinese microblogs with psychological lexicons. *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl*

Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, pages 844–849, 2014.

[65] A. Intxaurrondo, M. Surdeanu, O. L. De Lacalle, and E. Agirre. Removing noisy mentions for distant supervision. *Procesamiento del lenguaje natural*, 51:41–48, 2013.

[66] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9(6):1–12, 06 2014.

[67] M. Jaitner. Exercising power in social media. *The fog of cyber defence*, page 57, 2013.

[68] Y. Jiang and J. Jiang. Understanding social networks from a multiagent perspective. *IEEE Transactions on Parallel and Distributed Systems*, 25(10):2743–2759, Oct 2014.

[69] D. Jing-wei, D. Kai-ying, L. Yong-sheng, and L. Ying-xing. Study on evolution model and simulation based on social networks. In *2012 Eighth International Conference on Natural Computation*, pages 1238–1241, May 2012.

[70] P. K. Jonason, A. Jones, and M. Lyons. Creatures of the night: Chronotypes and the dark triad traits. *Personality and Individual Differences*, 55(5):538 – 541, 2013.

[71] E. C.-C. Kao, C.-C. Liu, T.-H. Yang, C.-T. Hsieh, and V.-W. Soo. Towards text-based emotion detection a survey and possible improvements. In *Information Management and Engineering, 2009. ICIME'09. International Conference on*, pages 70–74. IEEE, 2009.

[72] G. King, J. Pan, and M. E. Roberts. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501, 2017.

[73] S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[74] B. Kirman, C. Lineham, and S. Lawson. Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 121–130. ACM, 2012.

[75] M. Klier, J. Heidemann, M. Gneiser, and C. Weiá. Valuation of online social networks - an economic model and its application using the case of xing.com. In S. Newell, E. A. Whitley, N. Pouloudi, J. Wareham, and L. Mathiassen, editors, *ECIS Proceedings*, page paper 442, 2009.

[76] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proc. 14th International Conference on Machine Learning, 1997*, 1997.

[77] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[78] A. Kowcika, A. Gupta, K. Sondhi, N. Shivhre, and R. Kumar. Sentiment analysis for social media. *International journal of advanced research in computer science and software engineering*, 2013.

[79] S. Kumar, F. Spezzano, and V. Subrahmanian. Accurately detecting trolls in slashdot zoo via decluttering. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 188–195. IEEE, 2014.

[80] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750. ACM, 2009.

[81] R. Lambiotte, J. Delvenne, and M. Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, July 2014.

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

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

[84] A. Lökk and J. Hallman. Viability of sentiment analysis for troll detection on twitter: A comparative study between the naive bayes and maximum entropy algorithms, 2016.

[85] G. Lombardo, A. Ferrari, P. Fornacciari, M. Mordonini, L. Sani, and M. Tomaiuolo. Dynamics of emotions and relations in a facebook group of patients with hidradenitis suppurativa. pages 269–278, 2018.

[86] H. T. Maddali, P. A. Gloor, and P. A. Margolis. Comparing online community structure of patients of chronic diseases. *CoRR*, abs/1502.05263, 2015.

[87] G. Matrella, G. Parada, M. Mordonini, and S. Cagnoni. A video-based fall detector sensor well suited for a data-fusion approach. *Assistive Technology from Adapted Equipment to Inclusive Environments, Assistive Technology Research Series*, 25:327–331, 2009.

[88] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[89] T. Mihaylov, G. Georgiev, and P. Nakov. Finding opinion manipulation trolls in news community forums. In *CoNLL*, pages 310–314, 2015.

[90] T. Mihaylov and P. Nakov. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 399–405, 2016.

[91] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Pulse of the nation: Us mood throughout the day inferred from twitter. *Northeastern University),[online]*, 2010.

[92] T. Mitchell. Conditions for the equivalence of hierarchical and flat bayesian classifier. *Technical report, Center for Automated Learning and Discovery, Carnegie- Mellon University*, 1998.

[93] S. 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*, SemEval '12, pages 246–255, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[94] S. M. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement*, 2015.

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

[96] L. Morrissey. Trolling is a art: Towards a schematic classification of intention in internet trolling. Technical report, Griffith Working Papers in Pragmatics and Intercultural Communications, 3 (2), 2010.

[97] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter, 2013.

[98] A. Onderdijk, H. Van der Zee, S. Esmann, S. Lophaven, D. Dufour, G. Jemec, and J. Boer. Depression in patients with hidradenitis suppurativa. *Journal of the European Academy of Dermatology and Venereology*, 27(4):473–478, 2013.

[99] F. J. Ortega, J. A. Troyano, F. L. Cruz, C. G. Vallejo, and F. EnríQuez. Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks*, 56(12):2884–2895, 2012.

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

[101] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[102] W. G. Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[103] B. Pfaffenberger. " if i want it, it's ok": Usenet and the (outer) limits of free speech. *The Information Society*, 12(4):365–386, 1996.

[104] R. Plutchik and H. Kellerman. *Emotion: theory, research and experience*, volume 3. Academic press New York, 1986.

[105] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63, 2014.

[106] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics, 2012.

[107] F. Riquelme and P. González-Cantergiani. Measuring user influence on twitter: A survey. *Information Processing & Management*, 52(5):949–975, 2016.

[108] M. Roccetti, G. Marfia, P. Salomoni, C. Prandi, R. M. Zagari, F. L. G. Kengni, F. Bazzoli, and M. Montagnani. Attitudes of crohn's disease patients: Infodemiology case study and sentiment analysis of facebook and twitter posts. In *JMIR public health and surveillance*, 2017.

[109] B. Roth, T. Barth, M. Wiegand, and D. Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78. ACM, 2013.

[110] H. Saif, M. Fernández, Y. He, and H. Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

[111] L. Sani, M. Amoretti, E. Vicari, M. Mordonini, R. Pecori, A. Roli, M. Villani, S. Cagnoni, and R. Serra. Efficient search of relevant structures in complex systems. In G. Adorni, S. Cagnoni, M. Gori, and M. Maratea, editors, *AI*IA 2016 Advances in Artificial Intelligence*, pages 35–48, Cham, 2016. Springer International Publishing.

[112] L. Sani, G. Lombardo, R. Pecori, P. Fornacciari, M. Mordonini, and S. Cagnoni. Social relevance index for studying communities in a facebook group of patients. In K. Sim and P. Kaufmann, editors, *Applications of Evolutionary Computation*, pages 125–140, Cham, 2018. Springer International Publishing.

[113] C. W. Seah, H. L. Chieu, K. M. A. Chai, L.-N. Teow, and L. W. Yeong. Troll detection by domain-adapting sentiment analysis. In *Information Fusion (Fusion), 2015 18th International Conference on*, pages 792–799. IEEE, 2015.

[114] M. Sewell. Ensemble learning. *UCL Research Note*, 11(02):1–12, 2011.

[115] C. N. Silla Jr and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[116] N. F. F. Silva, E. R. Hruschka, and E. R. Hruschka. *Biocom Usp: Tweet Sentiment Analysis with Adaptive Boosting Ensemble*. University of Sao Paulo and Federal University of Sao Carlos (Sao Carlos,Brasil), 2014.

[117] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[118] C. Strapparava, A. Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[119] J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer, 2013.

[120] J. Synnott, A. Coulias, and M. Ioannou. Online trolling: The case of madeleine mccann. *Computers in Human Behavior*, 71:70–78, 2017.

[121] S. Tobi, S. Ma'on, and N. Ghazali. The use of online social networking and quality of life. In *2013 International Conference on Technology, Informatics, Management, Engineering, and Environment*, pages 131–135, June 2013.

[122] R. Ugolotti, F. Sassi, M. Mordonini, and S. Cagnoni. Multi-sensor system for detection and classification of human activities. *Journal of Ambient Intelligence and Humanized Computing*, 4(1):27–41, 2013.

[123] G. Wang, J. Hao, J. Ma, and H. Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1):223–230, 2011.

[124] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao. A depression detection model based on sentiment analysis in micro-blog social network. In J. Li, L. Cao, C. Wang, K. C. Tan, B. Liu, J. Pei, and V. S. Tseng, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, pages 201–213, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[125] J. L. S. Yan, H. R. Turtle, and E. D. Liddy. Emotweet-28: A fine-grained emotion corpus for sentiment analysis. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016, page 1149–1156, 2016.

[126] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, pages 412–420, 1997.

[127] A. Younus, M. A. Qureshi, M. Saeed, N. Touheed, C. O'Riordan, and G. Pasi. Election trolling: analyzing sentiment in tweets during pakistan elections 2013. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 411–412. ACM, 2014.

[128] X. Zhu, S. Kiritchenko, and S. Mohammad. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 443–447, 2014.