



UNIVERSITÀ DI PARMA

DOTTORATO DI RICERCA IN

Tecnologie dell'Informazione

Ciclo XXXI

Analysis of Dynamics and Credibility in Social Networks

Coordinatore:

Chiar.mo Prof. Marco Locatelli

Tutor:

Chiar.mo Prof. Michele Tomaiuolo

Dottorando: *Giulio Angiani*

Anni 2015/2018

*a Tiziana, Agnese e Agata
... in ordine di apparizione*

*"They won't listen. Do you know why?
Because they have certain fixed notions about the past.
Any change would be blasphemy in their eyes,
even if it were the truth.
They don't want the truth; they want their traditions."*

Isaac Asimov, Pebble in the Sky

Index

Abstract	1
Introduction	3
I Social Networks Dynamics	5
1 Models of Participation in Social Networks	7
1.1 Types of virtual communities	8
1.2 Requirements and features of Online Social Networks	9
1.3 Models of participation	10
1.3.1 Social Network Analysis	10
2 Information Spreading in Social Media	13
2.1 Social capital	14
2.2 Information spreading	16
2.2.1 Kinds of social networks structures	16
2.3 Motivations for participation	17
2.4 Anti-social behaviours and trolling	20
II Information investigation: A Machine Learning approach	23
3 Social Network Content Credibility	25

3.1	Recognizing false information	27
3.1.1	Types of false information	27
3.1.2	Impact of false information	28
3.1.3	Main characteristics of false information	30
3.1.4	Detection of false information	34
4	Social Network Content Analysis	39
4.1	Techniques for general-purpose classification	39
4.1.1	Classification	40
4.1.2	Regression	44
4.1.3	Clustering	46
4.2	Techniques for text analysis	48
4.2.1	Text preprocessing	48
4.2.2	The Bag-of-Words model	51
4.2.3	Text classification	53
4.3	Techniques for image analysis	58
4.3.1	Object recognition	58
4.3.2	Google Cloud Vision	59
4.3.3	Image manipulation recognition	67
4.4	Techniques for network analysis	69
4.4.1	Community detection	70
4.5	Conclusions	72
III	Credibility Discovery in Social Networks	73
5	Text based trustiness analysis	79
5.1	Data collecting and preprocessing	79
5.1.1	Public dataset	80
5.1.2	Self-made dataset	81
5.2	Data analysis and classification	82
5.3	Results and conclusions	83

Index	iii
6 Image based trustiness analysis	87
6.1 Data collecting and preprocessing	87
7 Social graph based trustiness analysis	91
7.1 Collecting data about social network users	91
7.1.1 Retrieving trusted and untrusted users	92
7.1.2 Initial credibility value	93
7.1.3 Reliability estimation	93
7.1.4 Reliability estimation by behavior	95
7.1.5 Classification	96
7.1.6 Credibility prediction by regression	98
8 Source structure based detection system	101
8.1 Collecting data about web sites	102
Bibliography	107
Acknowledgments	115

List of Figures

2.1	A complete graph	17
2.2	An hypercube based representation of social network	17
2.3	A random graph	18
3.1	Time between patrolling and flagging <i>(Reprinted with permission from [52])</i>	29
3.2	Number of pageviews per day <i>(Reprinted with permission from [52])</i>	30
3.3	Suspicion of fake reviews related to appreciation level <i>(Reprinted with permission from [52])</i>	32
3.4	Fraudulent reviewers often operate in a coordinated or “lock-step” manner, which can be represented as temporally coherent dense blocks in the underlying graph adjacency matrix <i>(Reprinted with permission from [52])</i>	33
3.5	An example of graph-based fake review detection	37
4.1	Standard steps in classification process	41
4.2	Levels of accuracy related to classifiers and number of features . . .	44
4.3	Regression applied for predicting end-of-May performances in Ital- ian subject	45
4.4	Regression applied for predicting end-of-May performances in Math subject	45

4.5	PCA 3D-representation of real students' final outcome	47
4.6	Clustering applied for predicting end-of-year outcome starting from first three months data	47
4.7	Generic text mining process, with focus on pre-processing techniques	49
4.8	Hierarchical classification structure. The six subjective categories cover the Parrott sentiments classification.	57
4.9	The example image posted to Google Cloud Vision system	59
4.10	GCV: face and face-emotion detection	60
4.11	GCV: web entities detection. This information comes from Google internal knowledge base. It includes data related to other sites where Google can find similar images or the same elements of the submitted picture.	60
4.12	GCV: OCR detection	61
4.13	GCV: Safe Search results	61
4.14	A famous example of fake news based on image manipulation . . .	68
4.15	Another famous fake image: the Sandy hurricane shark	68
4.16	Different kinds of relations between nodes	70
4.17	Social network graph	71
4.18	Communities detected in the same social network graph	71
4.19	The complete process pipeline	76
5.1	Text representation of an emoticon	81
5.2	Example of self-built textual dataset	82
5.3	Accuracy according different features number for DS1	83
5.4	Accuracy according different features number for DS2	84
5.5	Confusion matrix calculated on FakeNewsNet dataset DS1	85
5.6	Confusion matrix calculated on dataset DS2	85
7.1	different users groups descrimination	92
7.2	A graph representation of downloaded data	94
7.3	Reliability calculated by social graph connections	95
7.4	Reliability calculated for Facebook pages	96

List of images

vii

7.5	bipartite graph of users-pages interactions	97
7.6	Comparison between algorithms in classification	98
7.7	Regression estimation errors. About half predicted results differ less than 0.1 from the values calculated with Alg. 1	99
8.1	Web sites features data set example	103

Abstract

*You'll never find the truth
if you don't accept also things
that aren't in your expectations*

– Heraclito

According to the view of many researchers, today one of the greatest challenges on the net is the ability to discriminate the truth from falsehood, while moving along the mazy world of the Web and Online Social Networks. Many attempts have been made with this goal in mind; however, there are no shared solutions yet.

The aim of this thesis is to propose a methodology to tackle this problem, starting from an in-depth analysis of the main interactions that occur within Online Social Networks. In fact, the analysis starts with a description of the different ways of interaction in Online Social Networks, to better explain why people massively use these instruments today. The focus is mainly on social network dynamics, with particular attention to the models of participation and to the reasons that push people to be connected. It is also very important to understand how these dynamics help the dissemination of information. A relevant part of this work is related to describing the information spreading techniques and phenomena.

Then, the concept of credibility on the Web is furthermore investigated, with specific focus on Online Social Networks. Among the questions to be addressed, the most relevant ones are: What is true and what is false? Why do we trust some information instead of others? How much is our social network significant for trusting

or untrusting some news?

After explaining this social background and demonstrating how the social context can influence the perception of truth on the net, a model is proposed, with the aim to help users to estimate to what extent a piece of information can be considered true. The estimation is based on four aspects: *(i)* the credibility of the source that publishes it and the users who share it, including the reliability of the social relations of the source; *(ii)* the structure of the source site; *(iii)* the text used to spread the information, including the sentence structure and the used words; and *(iv)* the use of images.

All these aspects have been dealt with, using different machine learning techniques. At a first stage, each aspect has been analyzed independently of the others. These different modules lead all to very promising results. A further step of analysis, which is modeled in this work, requires a composite system to put all the results together.

Introduction

*“The cultured man is the one who knows
where to go looking for information in the only
moment of his life in which he needs”*

– Umberto Eco

The research work presented in this dissertation is essentially three-fold.

Firstly, in Part I, the existing dynamics in Online Social Networks (OSNs) are investigated, focusing on the different ways to participate inside them. Mainly four kinds of participation have been found: *(i)* individual motivations; *(ii)* relational capital; *(iii)* cognitive capital; and *(iv)* structural capital. All these motivations are analyzed, disserting also about the various types of online communities and about the various metrics used for understanding how a user is situated inside his/her own network (i.e. Degree Centrality, Betweenness Centrality, and Closeness Centrality). Furthermore, the processes of information spreading are explained. In fact, these processes can be modeled to represent the roles of different users, aiming at predicting the future impact of a new information released within the Online Social Network. Then, the focus moves on the analysis of social relationships, in the different forms which can be expressed, and on the concept of *information credibility*. Some techniques for representing the social relations of a user are described, for investigating in particular the user relationships supported by Facebook and Twitter. After collecting relationship data, the social network can be represented with a graph $G = (V, E)$, where V is a discrete finite set of nodes (or vertices) that represents the people or

users involved, and E is a binary relation on V , which represents relationships among users. The neighborhood of a node is the set of other nodes directly connected to him. The higher a node grade, the greater the importance of a user in the network. Using this structure, and the analytical operations on a graph, is possible to highlight the most influential people in a given community of users.

Then, Part II discusses the concept of information credibility: Are there any tools aimed at assigning a trustiness value to a news? When can a piece of information be considered trusted? This problem was identified as very important already in 2000 [25]. Misinformation, of course, was not born with the Internet, but, with the development of the new technologies, it has grown up tremendously. Many works, in these last years, have focused on this goal reaching different results: in this Part many of these approaches are shown for motivating the choice to develop a complex system for dealing with this problem from different points of view.

Finally, Part III of this dissertation is completely devoted to the structure of a newly proposed system for assessing the reliability of a piece of information (which could be a content, a user, a Web site, a Facebook page). This part is organized in five sections: one for each of the four subsystems for automatic credibility detection that have been devised and implemented according to a “monolithic” perspective, and a last section for giving some details about the final system – which uses as inputs the classifications of the other four subsystems. The four subsystems implement respectively four credibility detection processes, each one based on a different type of data associated with a piece of information: text; image; social graph; and finally source structure. The fifth section depicts the future work idea, which essentially consists of putting together the four subsystems, according to a “stacking” strategy.

Part I

Social Networks Dynamics

Chapter 1

Models of Participation in Social Networks

Social networks are around us

Social networking systems are bringing a growing number of acquaintances online, both in the private and working spheres. In businesses, several traditional information systems, such as those for Customer Relationship Management (CRMs) and Enterprise Resource Planning (ERPs), have also been modified in order to include social aspects. Social Network Analysis (SNA) can be useful to cope with common business problems, including: launching distributed teams, retaining people with vital knowledge for the organization, improving access to knowledge and spreading ideas and innovation. However, these goals are often frustrated by difficulties, including anti-social behaviours of participants, lack of incentives, organizational costs and risks. Participation in Online Social Networks (OSNs) has long been studied as a social phenomenon according to different theories. This chapter discusses the basic aspects of Social Network Analysis and some theories of participation in social networks, inspecting in particular the role of social capital.

1.1 Types of virtual communities

Social networks may be characterized by a great variety of communities, even if it is possible to recognize some common traits: (i) the lack of central authority, (ii) the temporary nature, and (iii) the importance of reputation and trust in opposition to the traditional communication and information systems.

In 1994 Mowshowitz defines a Virtual Organization (VO) as “*a temporary network of autonomous organizations that cooperate based on complementary competencies and connect their information systems to those of their partners via networks aiming at developing, making, and distributing products in cooperation*” [63].

Some years later a VO will be defined also as “*flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources*” [29].

The main characteristic of Virtual Organizations is of course the technical availability of tools for effective collaboration among people located in different places, but one of the most important features that have contributed to the development of VO has been the possibility to share a large number of resources, including documents, data, knowledge and tools among interested people [5].

The last 20 years have witnessed a continuous increase of use of VOs, especially in distributed production and delocalisation, in sharing industrial project, in information spreading, in global logistics. In a single word: globalization.

Starting from the concept of VO, it is possible to explain what is meant for Virtual Team (VT). According to Powell et al., a Virtual Team is a “*group of geographically, organizationally and/or time dispersed workers brought together by information and telecommunication technologies to accomplish one or more organizational tasks*” [71].

Virtual Teams can represent organizational structures within the context of VOs, but they can also come into existence in other situations; in fact, it is absolutely frequent to find VTs which have no hierarchical structure but which are composed only by free persons that work together on a common project, for example an open source software.

1.2 Requirements and features of Online Social Networks

In Online Social Networks there are at least three distinct functional elements: *(i)* profile management, *(ii)* social graph management, and *(iii)* content production and discussion. In fact it is impossible to describe the behavior of a user without classifying its profile parameters, without inspecting the collocation inside its network or without understanding its published contents (i.e. texts, images, sharings). According to these elements, it is possible to categorize the OSNs in three different classes: *(i)* networks where the content is the most important element, *(ii)* networks with focus on the relationships and the interactions among users, and *(iii)* networks where both cited elements have the same importance.

The first type includes blogging, micro-blogging and media sharing web sites, like Twitter or Instagram. In these OSNs the relationship feature, which are typical for a system of this kind, are usually not symmetric. The content consequently has a great importance for sharing and for user following.

The second type instead, is usually represented by professional and business-related OSN, for example LinkedIn, where profile building is one of the most important and sensitive matter a user must deal with. In this kind of SN, a user pays great attention to the profile content and to getting relations with other users because his/her own success in the OSN depends mainly from these features.

The OSNs that belong to the last group are usually based on a mutual relationship (also called *friendship*) and on contents sharing. In these networks users usually have a public part, with some information which are visible to everyone, and a private or semi-private part whose contents (images, posts, other info) can be accessed only by specified users. Currently the most popular OSN in this category is Facebook which counts more than 2 billion users.

One of the goals motivating the participation in online communities is the benefit of team work over solo work. Moreover, openness is important for participation, too. In fact, a closed environment can hardly reach the minimal dimension and variety required for activating the typical dynamics at the basis of the different theories taken into consideration by analysts, for explaining participation in OSNs. However,

in some OSNs anonymity may also have a value in some activities of Virtual Teams, apart from encouraging participation in general. For example, an anonymous brainstorm activity may help opening a conversation about trust and ground rules for online meetings or in sharing sentiments or pathologies in self-help groups.

1.3 Models of participation

The result of the interactions among users in a social networking system is an Online Social Network, i.e., a special case of the more general concept of social network. A social network is defined as a set or sets of actors and the relations defined on them [86]. Social networks are typically studied using Social Network Analysis, a discipline that focuses on the structural and topological features of the network. More recently, additional dimensions have been added to the traditional Social Network Analysis approach [60, 11, 67, 44].

1.3.1 Social Network Analysis

Social Network Analysis is the process of studying social networks and understanding the behaviours of their members. Graph theory provides the basic foundations for representing and studying a social network. In fact, each member of the network can be mapped onto a node of the graph and each relationship between two members onto an edge that connects two nodes. In real life, it is very common to find examples of social networks: groups of friends, a company's employees, contributors with different aims, etc. In fact, SNA is currently used in many research fields including anthropology, biology, economics, geography, information science, organizational studies, political science and social psychology.

The main goals of SNA are:

- To investigate behaviors of some network users;
- To identify user memberships and position into sub-communities;
- To find possible relationships among users;

- To discover changes in the network structure over time.

Different aspects are useful for investigating the behaviours of a participant in a social network: the most relevant are his/her position in the social network (i.e., which other members he/she is connected to) and his/her contributions to discussions or collaborations (knowing which groups he/she belongs is an important information). Another important aspect is the kind of activity performed by a user in his/her social network [51]. Mainly, a user can be identified as “*active*” (when he/she produces contents, sends videos and photos, comments posts of other users, reports original texts and documents) or “*passive*” (when he/she is only a consumer of other users’ contents, limiting himself to liking or unliking those contents).

A second aspect, which is important to focus in, is the relationship between two members of the network [36]. Discovering the type of relationship between two members, their reciprocal trust and their distance in the network, is a basic information used by SNA to speculate about information diffusion and users contamination.

Another significant application of SNA is to find subgroups composed by different users, i.e., to perform community detection [28]. Detecting the presence of a community allows analysts to recognize the paths followed by information for reaching the network users. There are three main metrics for assessing a user position:

- Degree Centrality;
- Betweenness Centrality;
- Closeness Centrality.

Degree Centrality is connected to the concept of graph-node degree and tells us the number of direct connections a node has. The higher the degree, the stronger the capability to spread information to other users is. Betweenness Centrality is a gauging of how much a user could be able to diffuse information from a community to another, especially if he/she belongs to many communities. A very interesting approach aims at identifying influential users on the basis of their activity level, comparing it with the activity and reactions of their followers/friends [51]. Finally, Closeness Centrality is a measurement connected to the concept of graph-path length. It provides information

about how far a user is from all the users of his community: the shorter this value is, the greater the possibility to reach all the participants of the network is, when he posts a content.

The last major aspect, which SNA concentrates in, is to discover the changes of a social network structure during time [1]. Studying the dynamics of a network allows analysts to detect persistent relationships, if they exist, and also to discover the lead users. Lead users play an important role in the network, since they have the best marks, according to the main centrality metrics mentioned before, and remain stable in the network for a long period. Studying network changes can also be useful in predicting users' real connections [84].

Chapter 2

Information Spreading in Social Media

After analyzing the structure of an OSN for understanding the various ways to build a social network, and after focusing on the different ways in which a user can self-position himself/herself inside them, it is very important to explain also “why” a user chooses to belong to a network. An important theoretical foundation for the analysis of participation in social networks is constituted by *social capital*.

According to Jacobs [48], who studied this phenomenon in real social network, social capital represents a person’s benefit due to his/her relations with other persons, including family, colleagues, friends and generic contacts.

This concept is absolutely essential in better understanding the dynamics occurring inside a large users community. In fact, a user could join a network for many different reasons but always for receiving a kind of gain in doing it. This gain could be social, financial, human, but surely everyone which joins an organization usually receives a profit.

2.1 Social capital

As said above, social capital is another kind of capital which can be achieved by a user in joining a community. It is usually possible to consider the financial capital, which includes machinery and raw materials, and the human capital, which includes the additional knowledge and skills obtained by being part of a community.

Moreover, the human capital is strictly connected with the social one [58, 16]. Social capital is typically studied: (i) by drawing a graph of connected people and their own resources, creating a connection between each user's resources and those of his closest contacts; or (ii) by analyzing social structures in their own right, and supposing that the network structure alone can be used to estimate some user's competitive advantage, at the social stance [31].

The size of the ego-centered social network is an important factor to estimate the social capital of one individual; however, the size alone does not provide enough information. According to Burt [13] social capital is related with the number of non-redundant contacts and not directly with the simple number of contacts. In fact, although information spreads rapidly among homogeneous, richly interconnected groups, Granovetter [37] argues that new ideas and opportunities are introduced in the groups by contacts with people from outside the group. In order to explain this phenomenon, Granovetter distinguishes among three types of ties: (i) strong ties, (ii) weak ties, and (iii) absent ties.

The *strong ties* usually link close collaborators, friends and families.

One of the most helpful positions, that a user can achieve in social network, is just to represent a *weak ties* between two communities. In this way, he/she could be the only contact-point between two groups of users, becoming the entry point for spreading (or not spreading) an information inside these groups.

Another method to identify the motivations which induce a user to keep in contact with a community is to consider the following theories [60]:

- **Self-interest.** According to this theory, people usually create links with other people and participate in activities which maximize the satisfaction of their

own personal projects. This kind of interest can be declined in various fields, for example economics or politics.

- **Mutual interest and collective action.** In this case, the theory studies the coordinated action of individuals in a team. Users join in community for trying to reach goals which would be unreachable by individual action [33]. Of course, users meet in groups pushed by common interests.
- **Homophily and proximity.** This theory is based on the principle that users join a community mainly for contacting people with similar interest or lifestyle. However, not rarely it is possible to find dissimilar individuals which end up getting ties thanks to common similar users.
- **Exchange and dependency.** Another motivation for creating groups in a social network is the need to have a place where sharing and finding resources about a specific topic (for example musicians, sport teams, etc.). This theory just explains the structure and the interactions about different components.
- **Co-evolution.** The main concept of this social network theory is to study how individuals cooperate and compete to access limited resources, and how a community creates links internally and towards external communities.
- **Contagion.** The contagion theory explains how some ideas or pieces of information can be spread (or limited) over a network. In fact, ties between individuals can be used for promoting the diffusion of a news. On the other hand, separating users permits to limit this diffusion.
- **Balance and transitivity.** Since macroscopic patterns originate from local structures of social networks, balance theories [45] start from the study of triads in a digraph, or a socio-matrix. The typical distributions of triads configurations in real social networks show that an individual's choices have a consistent tendency to be transitive.
- **Cognition.** Finally, this theory is based on the study of how a community grows up depending on users knowledge and collaboration. In this way, the decision

to form a community is directly connected with the possibility to increase the common information gain.

2.2 Information spreading

In Social Network Analysis, studying the process of information spreading is a critical topic. As a matter of fact, understanding the dynamics of information (or rumor) spread in social networks is very important for many different purposes, such as marketing campaigns, political influence, news diffusion and so on. The way a piece of information reaches people and how much time it takes to do it are examples of analysis of information spreading processes. They depend mainly on *(i)* network characteristics – topology, dynamism, sparsity, etc. –, *(ii)* the meaning of the information content, and *(iii)* the influence of the source of information. These processes are absolutely central in this study because, as it is demonstrated in the following chapters, the estimation of credibility of a news strictly depends on these three characteristics.

Several models have been in fact developed in order to study such a phenomenon, but there is no unique standard option, due to the heterogeneity of social networks [61], that range from real-world to OSNs, such as micro-blogging services or forums.

Therefore here these theories have applied only to the Online Social Network communities, focusing on how users can be influenced by their own relationships and by the sources they retrieve information from.

2.2.1 Kinds of social networks structures

Despite those diversities, social networks share common features that are taken as basis for the analysis. Technically, a network can be represented as a graph $G = (V, E)$, where V is a discrete finite set of nodes (or vertices) that represents the people or users involved, and E is a binary relation on V , that represents relationships among users. The neighborhood of a node is the set of other nodes directly connected to him/her.

Depending on networks, the topological characteristics of the graph change, and several models have been investigated to match the correct shape of a network. Ex-

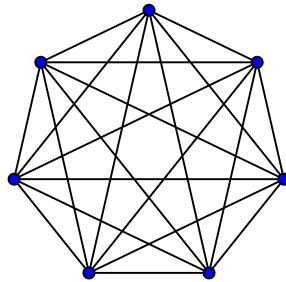


Figure 2.1: A complete graph

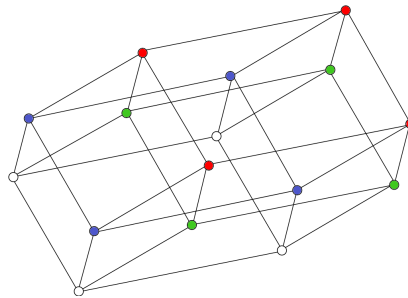


Figure 2.2: An hypercube based representation of social network

amples of such models are complete graphs [69] (like in 2.1), hypercubes [24] (like in 2.2), random graphs [21] and evolving random graphs [15] (like in 2.3), preferential attachment graphs [6, 18], power-law degree graphs [30] and so on.

There is no “best” model to represent a social network: it strictly depends on the specific network. The network studied in this work, which will be detailed later on, has been represented with a highly sparse graph, due to the great number of nodes (users) with a moderate number of relationships.

2.3 Motivations for participation

In order to understand the reasons that motivate the users in engaging in online social activities in general, and, more specifically, in sharing their valued knowledge

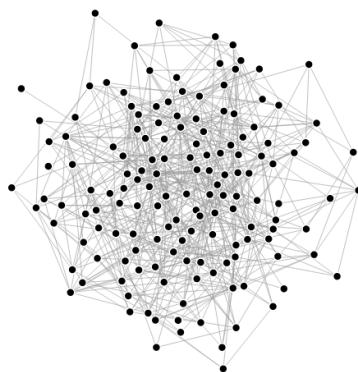


Figure 2.3: A random graph

in online communities, it is necessary to analyze (i) the nature and the structure of their relationships in the context of a specific community, and (ii) their implication over both online and offline reputation. Wasko & Faraj, for example, analyze the motivations for participation in a specific online Network of Practice [85]. In particular, the analyzed network is a public online forum of legal professionals, who participate under their real identities. The study takes the following features into account, as possible enablers of participation.

- **Individual motivations** One key aspect of social contribution is an individual's expectation that some new value will be created, as result of his participation in the network. The individual should expect to receive some benefits from his contribution, even in the absence of direct acquaintance with other members of the community and without mechanisms enforcing or encouraging reciprocity. Increasing the reputation is one of the most important forms of return of investment, especially if the online reputation is believed to have a positive impact on the professional reputation.
- **Relational capital** Another enabling factor for contributions to an online community is represented by the personal relationships among individuals, as members of that community. Relational capital is directly related to the level of an

individual's identification with the community, trust with other members [83], perception of obligation to participate and reciprocate, acceptance of common norms. In particular, commitment can be associated with a community, apart from individuals.

- **Cognitive capital.** Any meaningful interaction between two members of a community requires some basic shared understanding. All those common semantic resources, including languages, interpretations, narratives, contexts and norms, are usually described as cognitive capital. In fact, an individual can participate in community activities only if he/she possesses the required knowledge and, more in general, the required cognitive capital.
- **Structural capital.** Communities characterized by dense internal connections are dialectically correlated with collective actions. In fact, individuals who are strongly embedded in a social network, have many direct ties with other members and a habit of cooperation. On the other hand, an individual's position in the network influences his willingness to contribute, thus increasing both the number and quality of interactions.

Those factors have different weight in different social contexts. In the case study analyzed by Wasko & Faraj, reputation plays a crucial role, since it also affects professional reputation [85]. Other factors, though, also have significant correlation with the number and usefulness of contributions in the online community. The final results compare both the level and helpfulness of contributions against the following factors: (i) reputation, (ii) willingness to help, (iii) centrality in the network structure, (iv) self-rated expertise, (v) tenure in field, (vi) commitment, (vii) reciprocity. With regard to individual motivations, results for the case at hand show a stronger influence of reputation over intrinsic motivations, like willingness to help. Social capital, assessed by determining each individual's degree of centrality to the network, is confirmed to play the most significant role in knowledge exchange. Also cognitive capital, assessed by self-rated expertise and tenure in the field, shows a strong influence over participation, but this is mostly limited to the individual's experience in the field,

while self-rated expertise is not quite significant. Finally, in the analyzed Network of Practice, relational capital, assessed by commitment and reciprocity, is not strongly correlated with knowledge contribution, suggesting that these kinds of ties are more difficult to develop in an online network.

Both individuals and organizations also appreciate social media as they foster innovation, by improving collective thinking. In fact, creativity and innovation have long been notable subjects of organizational studies and social network analysis. Fedorowicz et al. [23] note that creative ideas rarely come from individuals. More often, they come from teams and groups, including those formed through social media. Dwyer [20] argues that, apart from the number of collaborators, it is also important to measure the quality of collaboration. In fact, various collaborator segments can be identified, with significant differences in the value of contributed ideas and the timing of participation. Thus, new metrics should be used, taking those differences into account and being based on information content. Hayne & Smith [41] note that groupware performance depends on the fit between the structure and task of the group. However, they argue that an important role may also be played by the cognitive structure, which also maps to the group structure. In fact, collaborative tasks may push human cognitive capabilities to their limits, in terms of perception, attention and memory. Thus, the authors argue for the integration of different areas of study, such as: psychology, especially with regard to abilities and limitations; theories of social interactions, with regard to group communication and motivation; analysis of groupware structures and human interactions mediated by artifacts.

2.4 Anti-social behaviours and trolling

In Computer-Mediated Communication (CMC), user behavior is very different from a face-to-face communication and every type of communication medium creates its own communication rules. Depending on the kind of CMC, users are allowed to variously adjust the degree of identity they reveal. The level of anonymity usually guaran-

teed in online discussions allows users to engage in behaviors they would otherwise be averse to carry out in face-to-face discussion. This lack of identity has contributed to the codification of new communication behaviors, like trolling [62].

Trolls are often seen as corrupters within an online community. They often share a group's common interests and try to pass as a legitimate participants of the group [19]. After that, they try to lead the conversation toward pointless discussion [43]. Morrissey [62] suggests that *"trolling is an utterer producing an intentionally false or incorrect utterance with high-order intention to elicit from recipient a particular response, generally negative or violent."*

A troll can damage a group in many ways. He can interrupt discussions, give bad advice, or undermine the mutual confidence of the user community. Trolls usually post a message into different sections (Cross-Posting), by doing this they are able to annoy more groups simultaneously. Nowadays many companies are using tools such as blogs, forums, social media (including self-developed ones) for their own interests. Trolls are therefore a threat to private social platforms as well as for public ones. The most widely used solution against trolls is to ignore provocations. Some systems provide filters (killfile, blacklist) that allow to exclude trolls from public discussions.

The most widely used solution against trolls is to ignore provocations. Some systems provide filters (killfile, blacklist) that allow to exclude trolls from public discussions. In recent years, many projects have been developed for the automatic detection of trolls in online communities. Some works [75] use a supervised learning algorithm, which allows to classify the polarity of posts and identify trolls as users with a high number of negative messages. The classifiers are trained using examples of positive and negative sentences. The polarity classifier is trained on a data set of movie reviews written in standard English. The Support Vector Machine algorithm is used to do binary classification of trolls. Since the data set contains messages from different topics (different forums), some domain adaptation techniques are used to get better results.

Furthermore, the frequency of messages, and possibly also the frequency of generated answers, is another factor for determining the presence of a troll in the network: the higher the frequency, the higher the probability that he is a troll [12]. Ortega et al.

[65] propose a method to compute a ranking of the users in a social network, regarding their reliability. The goal is to prevent malicious users to gain a good reputation in the network. To achieve this purpose, they create a graph taking the users of the network as the nodes. The edges represent the opinions of some users about others, and the weights of the edges correspond to the intensity of the relationship between the nodes. Galán-García et al. [34] the authors suppose that “it is possible to link a trolling account to the corresponding real profile of the user behind the fake account, analysing different features present in the profile, connections data and tweets characteristics, including text, using machine learning algorithms.” In fact, machine learning techniques can be used to associate users’ posts with various emotions, in addition to generic positive or negative sentiments [26, 27].

More recently, researchers from Stanford and Cornell Universities have developed an algorithm that can estimate the need to ban a member of an online community, after observing only five to ten online posts [14]. In particular, the authors present a data-driven approach to detect antisocial behavior in online discussion. The data sets are collected from users that have been banned from a community.

Part II

Information investigation: A Machine Learning approach

Chapter 3

Social Network Content Credibility

As pointed out in the previous chapter, the credibility of information, especially in social networks, has become a very high matter. This section better investigates the different ways used for spreading false information on the Web.

To simplify the explanation, one can divide the false information considering the intent of spreading and its content. Furthermore, it is possible to only focus on the information that is widely accessible and that can be received by a large number of users at the same time (for example fake tweets, false reviews, altered images).

On the basis of the intent, false information can be identified as *misinformation*, created only for disseminating false news or data, and *disinformation*, created instead with the intent to damage the target of the information itself.

Both can have negative impacts on users, but the second one must be considered more dangerous because the main goal is clearly to hurt the target with negative content.

On the basis of the content, false information can be divided into two groups: *opinion-based* and *fact-based*.

The opinion-based information consists of all the news which do not have a real ground truth but come from personal ideas about a person or a product (for example

a review on an e-commerce site or a post about a politician). On the contrary, a fact-based information could be built with other false information (for example a post linked to a page of a website that contains only completely made up news).

After generating false information, it must obviously be spread on the Web: to do this, different methods are used, *fake accounts*, *bots*, *sockpuppets*. These instruments can be created and controlled by the same user or by different users, with the common aim to advertise the target information as more as possible.

This dissemination operation usually permits to reach many users with the same message and, at the same time, to create a false social consensus around a false piece of information, e.g. by retweeting or reposting in many accounts.

In order to discriminate true users from bots or sockpuppets, it is very important to understand their position in the social network (the *centrality* that is investigated in subsection 1.3.1).

Chapter 7 reports a discussion on how this kind of information could be used to estimate the credibility of a user using only data about its position inside its own social network and about its interactions with different contents.

However, the spreading of false information could be blocked if readers were able to identify it. Actually the real problem is that humans can understand if an information is true or false only with accuracies between 53% and 78%, according to different experiments. Moreover this percentage decreases deeply especially for two reasons: if a user agrees with the content that occur in a post and if the piece of information is well written and not too short.

This human characteristic highlights exactly the problem which this project tries to solve: providing a technological support to identify, or at least, to put in evidence a content on a page, estimating automatically its credibility value.

The next sections describe the techniques actually used for recognizing false information and then the technologies used to analyze texts, images, and social network in general.

3.1 Recognizing false information

In many published works, the recognition of false information is approached by studying many aspects including the writing-style, the temporal features, the user properties, the user network properties, the spreading time. In several cases, it is possible to identify that real information has often very different values for these features: for example, in fake reviews, the text is on average longer than in real reviews, the window-time of the reviewer is usually shorter and related to a very small period, the false reviewers are often related, as they tend to cite each other.

3.1.1 Types of false information

As pointed out, false information can be categorized according to the content and to the intent. Now let us analyze this kind of classification with more detail.

Categorization based on content

False information can be categorized according to content as *opinion-based* and *fact-based* [82]. In the first case, an author can create, more or less unconsciously, false information about the topic. In fact, expressing personal opinion about a specific target, he/she spreads new data which are not supported by proved facts. Clear examples of this kind of false information can be a fake review about a product published on an e-commerce site, but also a newspaper article written about a government act. The author conceals false information behind a personal opinion of the facts.

The aim of this behavior is usually to influence the readers and to convince them to change their ideas about the target of the fake piece of information. Furthermore, the author should use a text-style which is very similar to that present in real reviews or in real newspaper articles [70].

This is a very difficult challenge and represents a weakness in false information generation. This weakness is in fact very used for detecting automatically fake information [17]

Categorization based on intent

The intent of the author of a post can be classified as *misinformation* or *disinformation* [22] [42]. In the first case, the intent of the author is to spread information which is not verified or is only the result of the author's perception of reality, often based on his own experience (e.g., the author's idea about a new law or a new product could cause him to write "*The battery of that laptop is absolutely unuseful*" or "*The hall of that hotel was terribly dirty*"). The piece of information could be not real and could lead the readers to change their ideas about the target of the post. Differently, the disinformation is created with the precise intent to fool the readers about something or someone [70]. The most frequent example that can be found on the Web (but also in social messaging networks) is represented by posts built putting together true images of politicians or immigrants or religious leaders and sentences, which can be totally invented or real but told by other persons [47] [2]. Another great difference between these two kinds of false information is that the second one needs to be spread as more as possible to fulfill its real aim, i.e., misleading the readers. For this reason, usually, it is spread in relative short time and by large groups of users or bots or sock-puppets. This property can be used to identify a possible fake information.

3.1.2 Impact of false information

After investigating the definition of false information it is possible to focus on its impact first in real life and then in social networks. This impact must be analyzed separately, because there are very different characteristics between the two cases. In fact, in real world, disinformation can be measured usually looking at stock market value, analyzing how people react to natural disaster or terroristic activities. Instead, in the virtual world the impact can be measured in many ways according to the social network (e.g. Facebook, Twitter or Wikipedia) where this phenomenon is studied [32] [53] [79].

One of the most significant fact of hoax diffusion in social networks, is the time needed for reaching readers before being flagged as false news.

In [77] the authors show the average time spent before identifying an hoax on

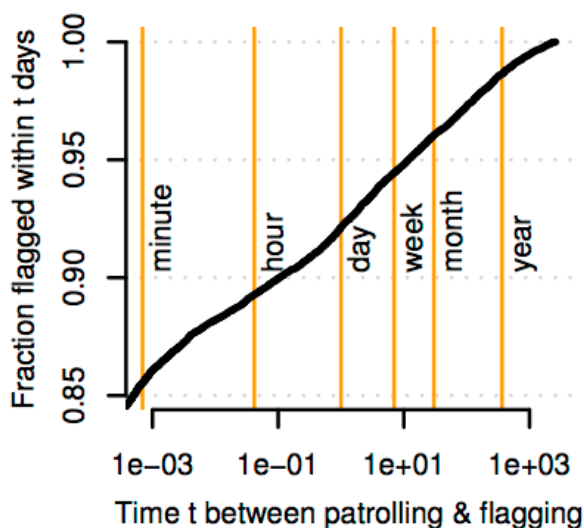


Figure 3.1: Time between patrolling and flagging
(Reprinted with permission from [52])

the Web is about 12 hours. Unfortunately, in the virtual world, this time is enough to reach a very large diffusion.

Figure 3.1 and 3.2 explain that the first hours are by far the most important for hoaxes and non-hoaxes diffusion.

As for the diffusion, hoax spread is measured in terms of number of related links, present in the web pages, that are clicked by users for reaching the false news. In [87] the authors, analyzing data on Facebook related to US 2016 President Election, focus on the diffusion of fake stories compared to diffusion of true stories highlighting that the first had had a sharing level significantly higher than the second (about 8700K against 7300K). Very similar results have been reached by Gupta et al. in [38], analyzing the spreading of news related to the Sandy hurricane of 2012.

Another very interesting work in this field is the paper of Frigerri et al. [32], where the authors analyze more than 4K rumors from the well-known site *snopes.com* which describes itself as an "Urban Legends Reference Pages" site.

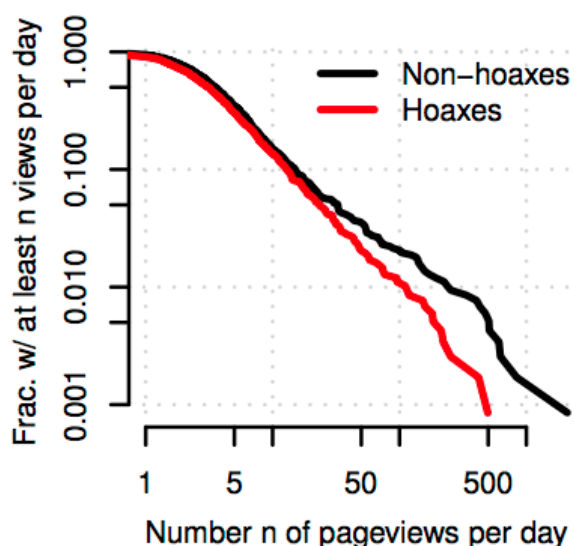


Figure 3.2: Number of pageviews per day (Reprinted with permission from [52])

After collecting these news, their dataset was composed by 45% of completely false stories, by 26% of real news and the other part by news with different level of truth. Then they have analyzed the propagation and the use of images and texts into the Facebook pages and have found that the diffusion of fake news was on average deeper than the one observed for real news. This study highlights also the large reach of false information on social media, fueled by its highly contagious nature.

3.1.3 Main characteristics of false information

Let us start to focus on the main characteristics that can be investigated to understand if a piece of information is true or false. Also in this case, it is possible to agree with [52] in categorizing these features in two groups, i.e., opinion and the fact characteristics. In both cases they are related to text, user, graph, rating score, time, and propagation. Part III explains how these features have been used to build different systems for estimating the truth level of an information.

Opinion-based characteristics

Below, the features most frequently used by researchers in the information truth estimation field of study are discussed.

Textual features. Since most posts, reviews, and other kinds of news which are spread on the Web, include texts, some of the most studied features for assigning a level of authenticity to information are those related to text. One of the first study on textual style of reviews [50] focuses on many Amazon reviews, calculating the similarity between them and highlighting that, in general, this similarity is very high only if calculated in reviews made from the same user. Other studies have also focused on linguistic characteristic of reviews, as number of words, average sentence length, used characters, number of verbs or adjectives, use of emoticons or also the expressed sentiment. Some results show that false reviews, especially negative ones, are generally shorter than real ones, have a very strong writing style (i.e. there are many strong adjectives like terrible or ugly, many punctuation signs like '!!!', more frequent use of capital letters).

In other works, Harris [40] demonstrated that false negative opinion are on average less readable than true negative reviews (measured by Average Readability Index [80]), and they are also more polarized regarding to the expressed sentiment.

Rating features Almost all the sites that allow users to leave a review about a product, or a service, usually allows the reviewer to express his/her level of appreciation with a 1-5 scale, often just selecting a star on a bar with five stars.

Many works investigated this particular features exploiting that, fake reviews, have usually a more polarized evaluation than the real reviews. In fact, in the false reviews, the greatest part of liking level is concentrated on level 1 and level 5 differently from the real reviews that have a more homogeneous distribution of evaluation [76]. Fig. 3.3 shows the result of a study about the review rate of real and fake reviews.

Temporal features From a temporal point of view, the most significant studied features are related to the interval between two consecutive reviews of the same users.

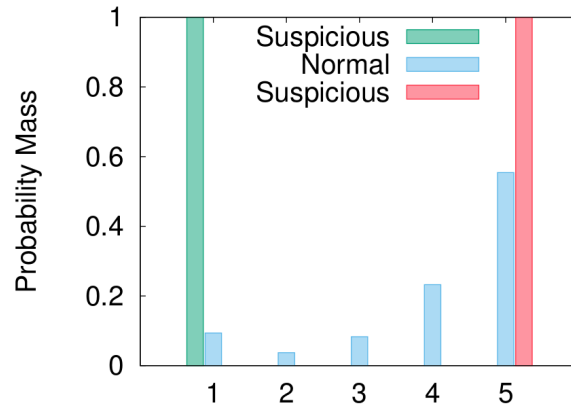


Figure 3.3: Suspicion of fake reviews related to appreciation level
(Reprinted with permission from [52])

Many papers in fact, demonstrate that usually users who create more reviews or posts, can be categorized in two groups: users that wait a significative time before writing another review and users that write more than a review in a short time [55]. Generally, the reviewers who posts fake news produce content with a very high frequency and usually in short time bursts.

Graph-based features Finally, analyzing the existing connections between authors and contents allows to represent the connections in graph-mode. In [10] another research finds a fraudulent pattern analyzing likes on facebook pages. In Fig. 3.4, it is possible to see a typical relation between users, pages and time. The figure explains in fact, how users that belong to a well-connected group like the same set of pages in a short defined time.

Fact-based characteristics

As pointed out, there is a great difference between opinions and fact. This section discusses about contents (reviews, posts , etc.) which can be only true or false. The focus is thus on hoaxes, rumors, fake news.

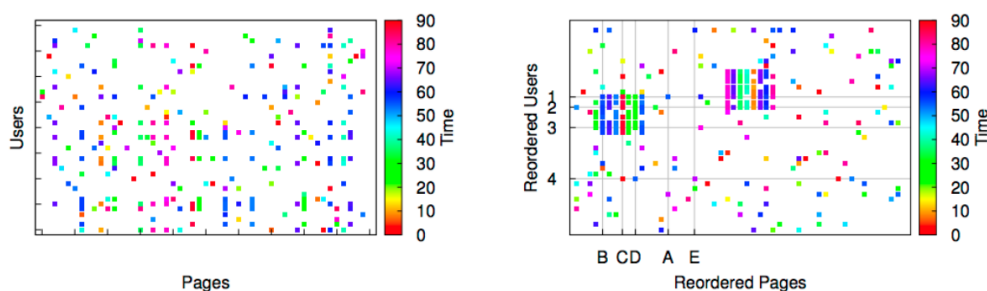


Figure 3.4: Fraudulent reviewers often operate in a coordinated or “lock-step” manner, which can be represented as temporally coherent dense blocks in the underlying graph adjacency matrix

(Reprinted with permission from [52])

Textual features The first analysis is related to textual characteristics. In many recent papers some textual features of hoaxes have been analyzed. Two articles titles are shown below, one of them is a fake.

1. BREAKING NEWS: World Fastest Man Usain Bolt In Critical Condition After Serious Car Accident In London. England
2. Typhoon Mangkhut: South China battered by deadly storm

Just looking, it’s possible to identify a great difference in writing style.

Fake news are usually longer than real news, tend to summarize all the information in the title, use many capital letters and punctuation to highlight their importance. Very often hoaxes can be confused with satirical news, as both use the same structure.

Focusing on the body content, it is possible to verify that fake news are shorter, repetitive and have fewer names or analytical words. They are simpler to read and do not contain technical specifications. This leads the reader to focus his attention only to the title, which could be also related to a topic different from the body [73].

Other researchers [53] found that hoaxes have similar textual properties as rumors and that they are surprisingly longer compared to non-hoax articles. In addition, they

typically contain far fewer web and internal Wikipedia references. "This indicated that hoaxsters tried to give more information to appear more genuine, though they did not have sufficient sources to substantiate their claims" [52].

User-related features Other important characteristics that must be studied for understanding the trustiness of a news, are those related to news creator. In many studies, like [73] [8], authors show that fake news creators have usually more recently registered accounts and have less editing experience. Furthermore, these creators are often bots which spread content in short-bursts of time [77].

Network features Relating to network position, hoaxes and fake news can be found focusing on the number of connections with other pages. It means that, as it seems obvious, if a news is real it is usually strongly connected with other similar pages and well-referenced to support the author in explaining his position. On the other hand, hoaxes are usually less connected than real news, and less supported by external sources. To overcome this behavior, hoaxes creators and/or bots are strictly connected to legitimate each other.

3.1.4 Detection of false information

The previous section has discussed many common characteristics that can usually be found in a false information spread on the Web, both for fact-based and for opinion-based ones. This section, instead, focuses on some techniques experimented by many researchers for automatically detecting false information. In several studies, many algorithms have been analyzed and these can be divided mainly into two groups: feature-based and the graph-based algorithms. The first group analyze the features summarized in the previous section for developing classification and regression algorithms. Algorithms belonging to the second one are focused on the relations between the creator of a fake information (or on the fake information itself) and other elements of the Web (i.e. other users of the same social network, linked sites, other pages or posts).

Let us start to analyze separately how fake information detection has been addressed by researchers. Part III, explains the model for facing up the problem. The model takes inspiration from both the already studied models, but put them together in an ensemble-learning system.

Feature-based detection

The first presented researches are related to the text-based detection, that is obviously connected with text features. Also because the text-related features are the most studied in this field given that one of the main characteristic of an information is just its text.

A logistic regression model to detect fraudulent reviews, using rating, textual features (i.e review title, body length, expressed sentiment, cosine similarity between review and product texts), and others, achieved an AUC of 78% [50].

In [64] instead, a Bayesian model has been built using cosine similarity between users and between reviews in addition to temporal features (see 3.1.3) and reviews rating. In this case the reached accuracy was 86%.

Relating to text-syntax, a useful work is [74] where the authors investigated both review syntax and semantic similarity. Syntactic similarity was measured using part-of-speech tags and semantic similarity using word-to-word distances in the WordNet synonyms database. This model, which has an F-score which varies between 0.5 and 0.7, but highlights well that many fraudulent reviewers use to post the same news just substituting some words with synonyms.

In [66] the authors have collected and analyzed 400 truthful and 400 positive-sentiment deceptive news from AMT¹-sourced reviews and trained a Support Vector Machine (SVM) classifiers using a variety of feature sets, such as n-grams and LIWC² features.

In [56] the authors apply an n -gram features based text-analysis to both customer reviews and deceptive employees reviews. In this work, using an SVM classifier and a semi-supervised model, they reach 65% accuracy.

¹Amazon Mechanical Turk <https://www.mturk.com/>

²Linguistic Inquiry and Word Count <http://www.liwc.net>

Graph-based detection

A completely different approach for false information detection is the analysis of relations which a piece of information has with other elements of the Net (i.e., users, pages, posts, etc.). Already in the first years of Internet, the existing relations between different kinds of elements presents on the Web have been studied using graphs [86], as this type of representation allows to better highlight some important characteristics of these elements (as it is explained in 1.3.1).

In this paragraph some approaches to graph-based fake news detection are described together with their main differences with the model proposed in this work.

First of all, a very interesting approach consists of following the propagation of rating in a graph where both users and products are present (i.e., graph nodes can represent both users and products). Each user U is connected to a product P if there is an evaluation E of the user about the product P . Each rating R_j can be measured as

$$R_j = \frac{1}{N} \sum_{i=0}^N E_{ij}$$

where N is the number of users and E_{ij} the rating of user U_i on product P_j and a corresponding graph is in Fig. 3.5

As in [72], the positive reviews created by “granted” users (i.e. the green-bordered users) are considered really a good rating and their negative reviews are considered a bad rating. Conversely, the negative reviews produced by unreliable users (i.e. the red-bordered ones) are considered like a positive rating to a certain product.

In the same work the authors use also this graph to follow the rating diffusion using also the notion of *homophily* (see section 2.1) which suggests that most honest users give genuine positive ratings to good products and negative ratings to bad products, and vice-versa for bad fraudulent users.

Other graph-based approaches have also been developed to identify fraudulent nodes in review networks, using different techniques, including edge distributions [46], dense block detection [49] and co-clustering [9]. This problem is very similar to identifying fake reviews or false information in general, as the intuition is that

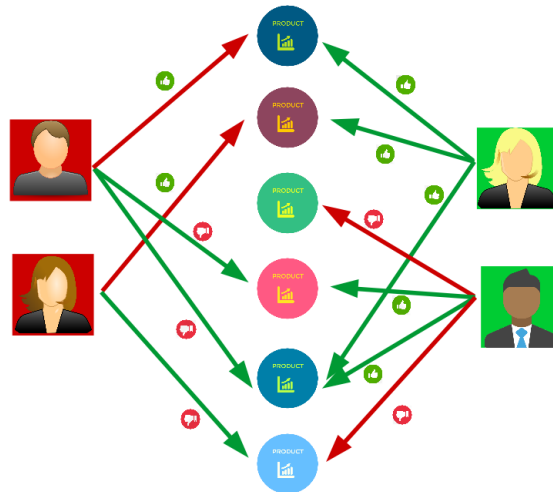


Figure 3.5: An example of graph-based fake review detection

by identifying fraudulent users, one can identify all their reviews or posts and then eliminate fake content from sites or social networks.

Chapter 4

Social Network Content Analysis

After deeply investigating the various motivations which encourage users to be on the Web and after better describing many different approaches for false information detection, this chapter analyzes in detail the techniques widely used for text, image and network analysis.

The last section briefly explains some classification and regression algorithms which have been used in this research. The chapter starts with a general introduction about the machine learning techniques, and then it focuses on those regarding with more detail text, image and network analysis.

4.1 Techniques for general-purpose classification

This section focuses on general classification techniques that have been used in the work for mining information from collected data. In particular, the use of classification and regression algorithms (in chapters 5, 6 and 8) and clustering (in chapter 7) was investigated. In the last chapter (8.1), an ensemble learning system has been designed and is still being developed, to put together the outputs of all the other systems in order to obtain a final trustiness estimation of information.

Machine learning techniques can be divided in four main groups which are (i) supervised, (ii) unsupervised, (iii) semi-supervised and (iv) reinforcement-based. Al-

most all ML problems can be traced back to the same structure: for each problem, a well-defined set of features have to be defined and each problem instance must have a set of values corresponding to the selected features. If instances are already tagged with labels (i.e. the correct output) the learning activity is called *supervised*, otherwise it is called *unsupervised*. In this case the aim of the research is to discover unknown classes of instances (*clustering*).

Another kind of machine learning is reinforcement learning. In this case the learning system does not receive directly an information about the correct decision to take, but learns the right behavior according to the signal received by the external training system. The greater the value of the signal, the better the chosen behavior.

4.1.1 Classification

The first shown example is related to a classification system trained with student performance data, collected in the Italian secondary schools for predicting the final year outcome [4].

Required data

Data used for this research have been extracted from the electronic logbooks of 10 high schools, located in different parts of Italy. All data have been anonymized in accordance with the current Italian privacy laws [35]. All information are related to marks obtained by students in their school tests and to their class attendance. Also end period marks (italian school year is usually divided into two periods) and end of year outcomes have been extracted for each subject. Informations about marks and attendance have been used for training some classifiers aimed at predicting the final outcome.

In accordance with the flow-chart showed in Fig.4.1, the first step performed on collected data has been a pre-processing and to trash out the invalid instances.

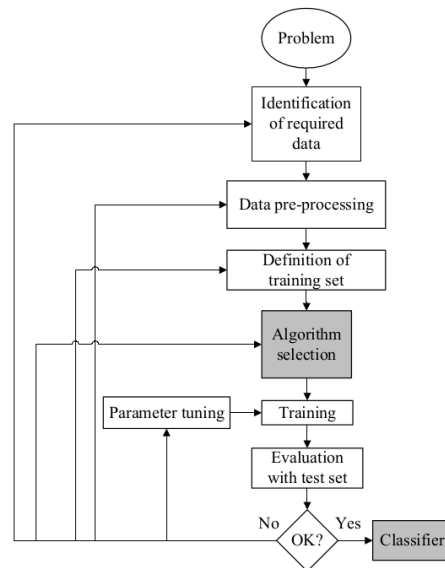


Figure 4.1: Standard steps in classification process

Data pre-processing

Since Italian high schools do not have standardized evaluation tests, marks are assigned by teachers in a 1-10 scale. Each teacher, for each test, can choose his/her way of evaluation. However, the assigned marks must belong to the [1-10] interval and are float values. Values outside this interval have not been used for our research. All the subjects have been clustered in six groups:

1. Italian Language (ita)
2. Mathematics (mat)
3. English Language (eng)
4. History (his)
5. Subjects strictly related to the student course (cou)
6. Other subjects (oth)

Feature selection and transformation.

Starting from daily raw-data, it has built the students' features, grouping data for each month with the following method: for each group, we calculated the average mark, collected in the period from Sep, 15th (start of school) till the end of each month, from October to May. The name assigned to these features has the following format, for values related to a single student: $\langle subject \rangle_ \langle month \rangle$; and the following format, for the average obtained using marks of all his/her classmates for the same period and for the same subject: $\langle subject \rangle_ \langle month \rangle_ grp$. The second group of features contains information about students school attendance and about the average attendance of his/her classmates. Each feature shows a student's school attendance in a certain period, as a number of days. Like in the previous case, these features have the following format: $abs_ \langle month \rangle$; and $abs_ avg_ \langle month \rangle_ grp$.

The third set of features is related to a student's trend in a certain period. For each student and for each subjects group, it has selected all marks related to that group, and it has calculated the linear regression line for these marks. The *trend* value is pointed out by the tuple (m, c, dev) ¹ which contains the values used to populate this third set.

Also in this case, the computed value has been compared with the corresponding average value of all the classmates. The same features have been calculated also for each group.

The last set of features contains only data about the school, the year, the course-year, the study course and about some data of end-of-school subjects marks. The final feature set F contains 410 elements.

Final dataset.

The whole dataset contains 13151 different instances, related to 10342 different students attending 10 Italian high schools. There are more instances than students, because for some schools we have collected data of several years. Each instance in-

¹ m is the slope and c is the y-intercept of the linear regression line, while dev is the standard deviation for the selected marks set

Table 4.1: Distribution of students' final results

Final result	Number of students	Percentage
POSITIVE	10609	80.7%
NEGATIVE	1100	8.6%
SUSPENDED	1442	11.0%

cludes 410 float value at most, one for each feature of the F set. Each instance has also the end-of-year outcome feature, which is mandatory in classification experiments. The outcome feature can assume one of the following values: “POSITIVE”, “NEGATIVE” or “SUSPENDED”². In Table 4.1, the distribution of students according to the final results is shown.

Since the best way to have a correct training is to have the same number of instances for each dataset class (i.e. dataset balancing), only 1000 instances for each categories have been used for the experiment, dividing them equally between the training and the test set.

Therefore both the definitive training set and test set are composed with 1500 instances (500 instances for each of the three classes).

Classifiers results

Many different classification algorithms have been applied on the dataset at hand obtaining sometimes very different results (Fig. 4.2 shows different levels of accuracy).

In particular the applied methods are a trained Neural Network, a Random Forest classifier, a Support Vector Machine algorithm (SVC) and the k-nearest neighbors algorithm (KNN).

The *Random Forest* classifier had better performance. Trying different algorithms and different sets of features allowed to identify the most significant characteristics for the studied problem.

²The “SUSPENDED” value indicates that the student must pass another exam at the end of August, for accessing the next class

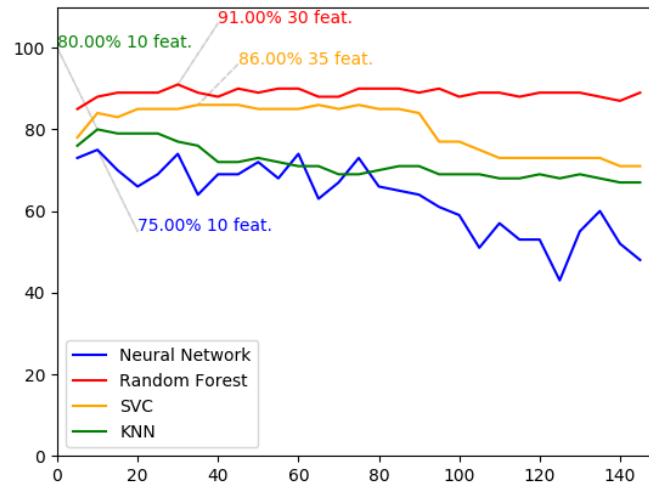


Figure 4.2: Levels of accuracy related to classifiers and number of features

4.1.2 Regression

Different from classification is *regression*. All the analyzed instances are represented by a set of values related to some features, but they are not labeled with a class name and the number of classes is potentially infinite.

With above specified data, an example of regression could be to predict the students' evaluations average in Italian and Math subject, starting by their activities data related only to the school first three months.

Differently from classification problems, in regression algorithms it is possible to calculate a float value as system output. In Fig. 4.4 predicted end-of-may marks in Italian and Math subject are shown. In both the experiments, the prediction errors remain quite always in range $[-0.5, +0.5]$.

Table 4.2 shows information related to mean squared errors and variance calculated on predicted values.

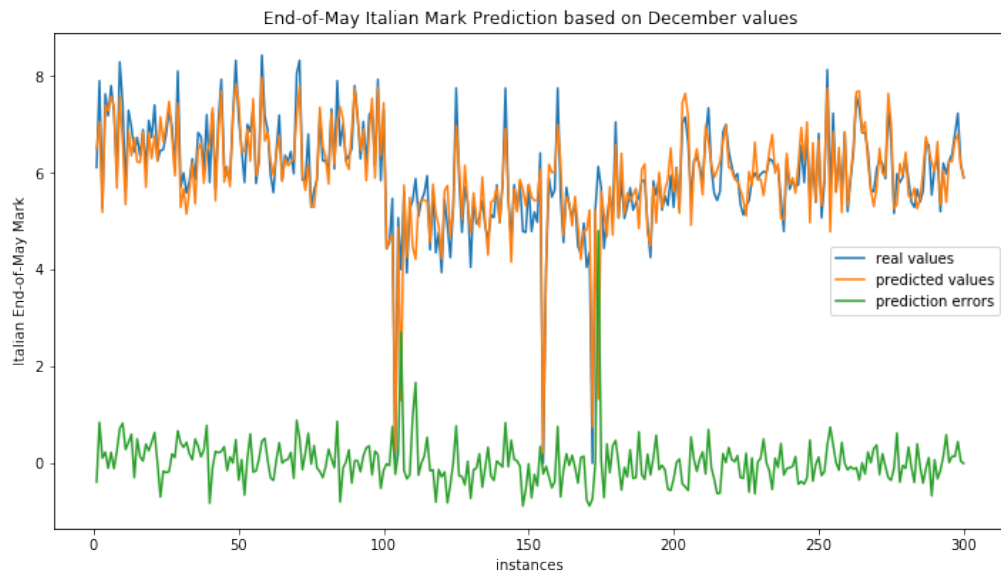


Figure 4.3: Regression applied for predicting end-of-May performances in Italian subject

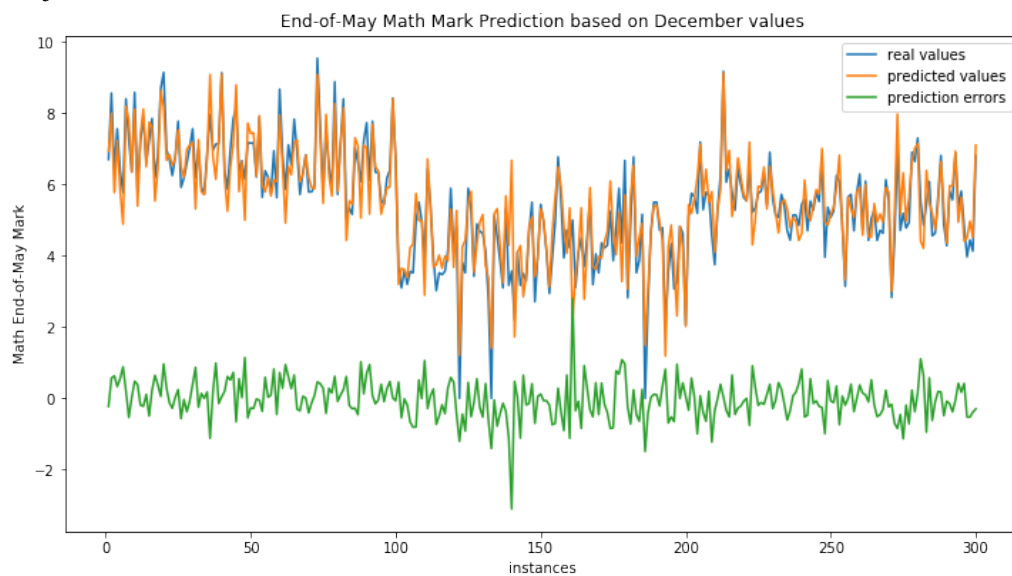


Figure 4.4: Regression applied for predicting end-of-May performances in Math subject

Table 4.2: Distribution of students' final results
Mean squared error Variance score

	Mean squared error	Variance score
Italian	0.24	0.79
Math	0.32	0.87

4.1.3 Clustering

The last technique that has been experimented is clustering, i.e., the ability to distribute a set of instances in different groups according to their represented characteristics. In the example, always taken by the same experiments treated before, each element of the analyzed set represent one student situation (i.e. all its marks and other information in school life) without an assigned class.

The K-means algorithm ³ has been used to divide all instances in K different groups taking together the most similar element: K-means is an iterative algorithm which, at each iteration, computes K *centroids* (i.e. the centers of K groups) and then calculates the *Euclidean distance* between all elements and the centroids. Then each element x_p is assigned to only one group according to the rule

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\}$$

After this assignment, the new *centroids* are calculated with the formula:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

It is guaranteed that the algorithm converges but not to find the optimal solution. The iteration stops when no more elements change groups.

In Fig. 4.6 the PCA 3D-representation of real clusters and predicted clusters is shown. In our experiments we reached only a 39% of accuracy in automatic clustering. In fact, more than half instances have been assigned to a wrong group.

³https://en.wikipedia.org/wiki/K-means_clustering

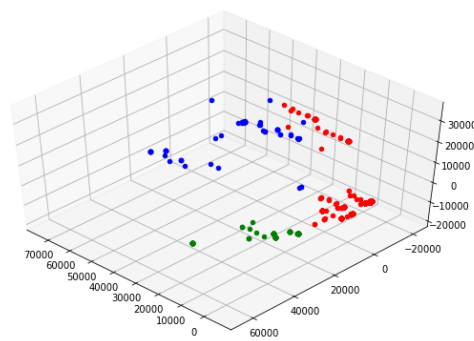


Figure 4.5: PCA 3D-representation of real students' final outcome

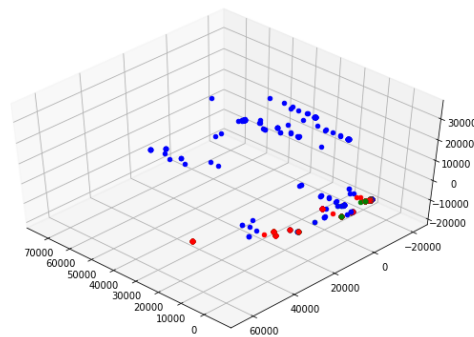


Figure 4.6: Clustering applied for predicting end-of-year outcome starting from first three months data

4.2 Techniques for text analysis

Text mining is the process of mining the useful information from text documents. It is an ensemble of techniques which allow the information extraction from both structured and unstructured texts, and it is widely used in many different fields of study, first of all the natural language processing (NLP), the text classification and clustering (TC), the sentiment analysis (SA).

In our research, as already pointed out, it is crucial to find those features which can better identify a fake information from a true one.

For all the analysis mentioned above, some steps are quite always used. This section starts to explain the main processes that, according with the literature, has been followed for this research.

4.2.1 Text preprocessing

One of the main peculiarities of NLP is certainly the fact that natural language is unstructured by definition. Each text written by a human is intrinsically full of different words, syntax, semantic, text-style, punctuation, use of grammar and, sometimes, also mistakes. Furthermore, every person can explain the same content with different methods, styles, opinions.

For all these reasons, a text can not be faced up for analysis like other data sources (i.e. reports, numbers, collections of structured data) but it needs to pre-process a textual input to make it structured and ready to be processed automatically.

In Fig. 4.7 three of the most common operations made in text preprocessing are shown.

Extraction

In the *extraction* step, the whole text to be analyzed is tokenized into singular words. This operation allow to represent any possible text as a set of words (sometimes repeated). With this representation the *bag-of-words* model 4.2.2 can be applied for mining.

An example of tokenization is to divide the simple sentence

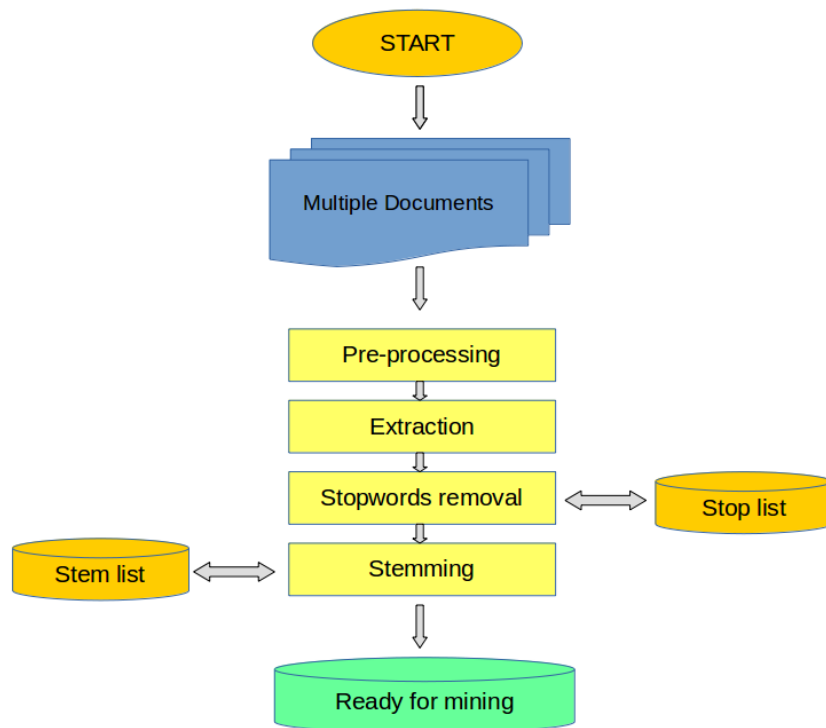


Figure 4.7: Generic text mining process, with focus on pre-processing techniques

"There is nothing either good or bad,
but thinking makes it so" ⁴

into the following list of words, without any further changes in words or letters.

```
['There', 'is', 'nothing', 'either', 'good', 'or',  
'bad,', 'but', 'thinking', 'makes', 'it', 'so']
```

Stop-words removal

However, before mining information from text, a good practice is to remove all those common words which are used in any context (including conjunctions, prepositions, articles and common verbal forms as 'are', 'is', etc.) [39]. These words must be removed because they are not useful, especially in the task of classification and information extraction. Furthermore, this technique reduces the raw data and improves the system performance. The gold-rule is that "*all not useful information is not an information*".

For this operation researchers usually use well-know stop-lists provided for all languages ⁵.

After applying a phase of stopwords removal , the previous list of words should (hopefully) contains:

```
['good', 'thinking', 'either',  
'nothing', 'bad,', 'makes']
```

only few meaningful words.

Stemming and lemmatization

The third typical operation performed on a text during preprocessing is stemming and lemmatization. This step replaces all words with the same stem (i.e., the same root) to

⁴W. Shakespeare. "Hamlet". Act 2 Scene 2

⁵<https://www.ranks.nl/stopwords>

only one word. A practical example of stemming is the substitution of all the verbal forms of one verb with its root.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is \Rightarrow be

car, cars, car's, cars' \Rightarrow car

The result of this mapping of text will be something like:

the boy's cars are different colors \Rightarrow the boy car be
differ color

One of the most famous and used stemming algorithm is the Lovins Stemmer proposed in 1968 [59].

Even often used together and sometimes confused, stemming and lemmatization do not perform the same operation on text. In fact, *stemming* usually refers only to drop out the ends of word with the aim to group similar words, whereas *lemmatization* instead, usually refers to a vocabulary for a morphological analysis of words to remove endings and return the base form of a word (i.e. the *lemma*).

An example of how the results of the two operations can be different is the application of them to the word *saw*.

The stemming operation might return only the char *s*, whereas the lemmatization might refer to different solution depending on whether the word is considered a verb (in this case it might return *see*) or a noun (in this case the word would have no modifications).

4.2.2 The Bag-of-Words model

In this research the BOW model is used for text classification and sentiment analysis. This model is useful when many texts T_1, T_2, \dots, T_n have to be analyzed together. After

applying the previous steps to all texts, what you get is the corpus dictionary which consists of a whole set S containing all the words (not replicated) that occur in any of the analyzed documents.

Hence, it is possible to transform each text T_i in a boolean vector V_i where each position j is related to one different word S_j and the value V_{ij} is 1 if $S_j \in T_i$, 0 otherwise.

An example of corpus dictionary is the following.

Given the three sentences

1. it was the worst of times
2. it was the age of wisdom
3. it was the age of foolishness

the resulting reordered words set of words S becomes

['age', 'foolishness', 'the', 'times', 'was', 'wisdom', 'worst']

The corresponding vectors are then

1. [0, 0, 1, 1, 1, 1, 0, 1]
2. [1, 0, 1, 1, 1, 0, 1, 0]
3. [1, 1, 1, 1, 1, 0, 1, 0]

After building the BOW model, dealing with new documents is very simple, although it is possible that they might contain different words not yet present in the model. In this case any new word is discarded and not used by the model. In so doing all analyzed texts have the same dimensional representation (i.e. vectors of the same size).

The sentence

- in the middle age there was great wisdom

would be represented with the vector

- [1, 0, 0, 0, 1, 0, 1, 1, 0]

The words [*middle*, *great*, *there*] are ignored.

Word scoring models

Not all words have the same importance in a text. This might seem obvious for a human, less for an automatic actor. After building the BOW model, it is necessary to understand which are the most meaningful words in the text.

For this purpose the most used score-system is the TF/IDF⁶ indicator. It works analyzing how many times a word is contained inside a document and, relating to other texts, how many times it is present in the others.

In this score model not all words are equally important or deemed interesting. The scores have the effect of highlighting words that are distinct (contain useful information) in a given document.

- **Term Frequency:** reports the frequency of the word in the current document.
- **Inverse Document Frequency:** evaluates how rare the word is across documents (the rarer, the better).

Looking to the previous examples, the TDF/IF value of the word *age* should be low because it is present in all the sentences (i.e. it is not so significant for any texts), on the other hand the word *wisdom* might have a high TF/IDF value because it is present just in one sentence.

4.2.3 Text classification

Text classification, or topic classification, is a methodology to categorize texts into predefined classes. Many business applications are based on this and some companies are taking advantage of. For example, when a customer communicates with a company asking for support, an automatic system can redirect the customer to the correct department.

⁶Term Frequency – Inverse Document Frequency

A very interesting application of text classification is sentiment analysis (SA). The aim of SA is to understand the perception of users regarding products or persons or places based on their comments. In SA the defined classes are usually two, *positive* and *negative*, but sometimes it is possible to introduce also a *neutral* class. It is widely used by companies, politicians, and researchers to track users behavior in social network.

Sentiment analysis

A SA process has been used for understanding the polarity of tweets downloaded from the Twitter⁷ channel *#Brexit* in October, 2015. The same technique has been used to classify texts into the two classes *hoax* and *no-hoax*, which will be detailed in Chapter 5.

The #Brexit case

In 2016 in the United Kingdom there was the very famous referendum for choosing if *remain* in the European Union or if *leave* it⁸.

In that period, that matter was strongly debated on social networks, where many thousands of people explained their own thinking and their sentiment about the possibility to leave the EU. Twitter was one of the most used social network and many researchers decided to analyze the tweets texts for understanding what the Britains would like and to try to anticipate the final result.

Data retrieving

First data was downloaded, by using API written by Twitter, about 570000 worldwide geolocated tweets, 360000 users, from June 21th, 2016 at 6:00 pm to June 25th, 2016 at 1:00 pm. These tweets were subsequently filtered by selecting only those from UK

⁷Twitter inc. <https://www.twitter.com>

⁸*Remain* and *Leave* were the two opposite political fronts; the first to stay in the EU, the second for exiting from it.

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

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 56000 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 APIs to provide data about gender, so it was possible to label gender from users name. In table there is summarized the classification of the tweets according the gender e the number of tweets per user.

Preprocessing

As explained in 4.2.1, we applied some preprocessing techniques to improve the accuracy of the emotion detection system. 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. Also the list of *emoticons* were processed, in order to represent them as tags (e.g., :) → 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 we try to make the text more uniform, as having different words written in the same way, helps 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.

We performed a sentiment analysis study on this database. In addition to positive and negative sentiments, we assigned also specific emotions to each instance (post or comment). In particular, we referred to Parrott’s socio-psychological model [68], which classifies all human feelings into six major categories:

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

In particular, our 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.

Training data

Since data have been collected directly from Twitter, it is extremely probable that it contains 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 and one for no-emotion (i.e. neutral sentiment), 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. We finally used this refined data set for training the hierachical classifier (view fig. 4.8).

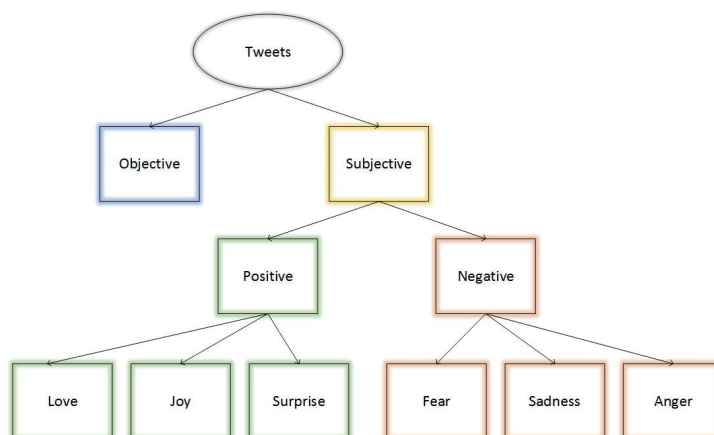


Figure 4.8: Hierarchical classification structure. The six subjective categories cover the Parrott sentiments classification.

Results

The tweets regarding Brexit have been divided into two timelines, before and after the release of the official referendum results, from June 21th, 2016 at 6:00 pm to June 23th, 2016 at 11:59 pm and from June 24th, 2016 at 00:00 am to June 25th, 2016 at 1:00 pm. In this way, it is possible to make a comparison and see indeed if there are changes in the obtained data. It has been able to properly represent the user thought with a single emotion. It is also interesting to underline that geolocated UK users that tweeted after June 24th, 2016 at 00:00 am are a number thirteen times higher (51927) than before ones (3867). Studying emotions only, some very interesting results can be observed, which deserve to be reported as maps of emotions. Furthermore, the analysis has checked whether the result of the referendum was predictable from the polarity obtained from tweets, considering also the influence of the Russian trolls.

As you can guess from Table 4.5, the most striking data is certainly the decrease of joy by almost 30.8%, but it is fascinating to notice that after the referendum sadness and fear have consistently increased by 12.5% and by 11% respectively.

Table 4.5: Percentage of emotions.

	Before Brexit	After Brexit
Anger	8.34 %	10.55 %
Fear	8.29 %	19.3 %
Joy	47.77 %	17.01 %
Love	2.06 %	3.49 %
Sadness	29.04 %	41.52 %
Surprise	4.5 %	8.13 %

4.3 Techniques for image analysis

This section shows the main techniques regarding images, with a particular attention to object recognition and to image manipulation detection techniques. The focus is on these two, as they are very central also for our definitive aim, the fake news detection.

4.3.1 Object recognition

In this work the image object recognition performed the task to investigate which kind of objects are contained in images connected to fake posts or fake news. One of the most used techniques for this aim are based on Convolutional Neural Network (CNN) or on Recurrent Convolutional Neural Network (RCNN) [57].

CNNs are a particular deep feed-forward neural network currently used in machine learning applications. Image analysis is one of those where this technique allows to reach very good results.

As the proposed work does not focus on CNNs or RCNNs, the theory of object recognition is not reported. In fact we used two external image analysis systems provided by Clarifai inc. ⁹ and by Google inc. ¹⁰.

⁹<https://clarifai.com/>

¹⁰<https://cloud.google.com/vision/> - Google Cloud Vision (GCV)



Figure 4.9: The example image posted to Google Cloud Vision system

Both the companies provide a service which processes a submitted image and returns a JSON structured response which contains all the information extracted from the image by the AI-system.

In particular we used GCV in the fake detection part, which is explained with detail in chapter 6.

4.3.2 Google Cloud Vision

It is shown here an example of use of GCV. During the experiments, a lot of images have been posted to the GCV system, in fig. 4.9, while in the 4.10, the most significant information is in a graphical form (for details see the Appendix ??).

Now it is explained with further details what kind of information is extracted from Google Cloud Vision API. The features, which were used in the project to estimate the trustiness of something related to an image 6, are: color composition in RGB, presence of faces and their relative emotions, the contained text, the objects tha occur

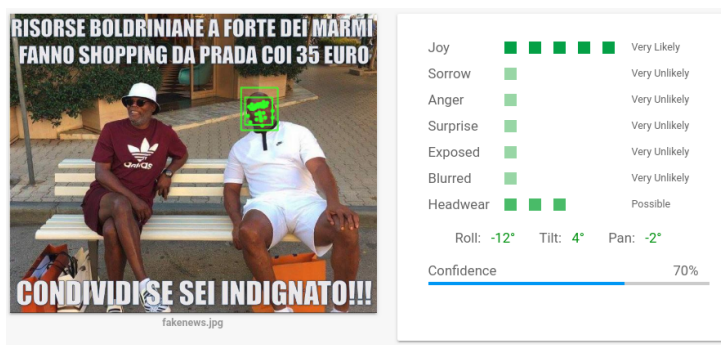


Figure 4.10: GCV: face and face-emotion detection

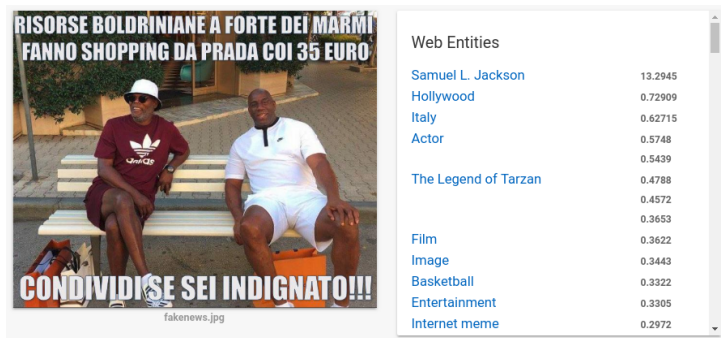


Figure 4.11: GCV: web entities detection. This information comes from Google internal knowledge base. It includes data related to other sites where Google can find similar images or the same elements of the submitted picture.



Figure 4.12: GCV: OCR detection

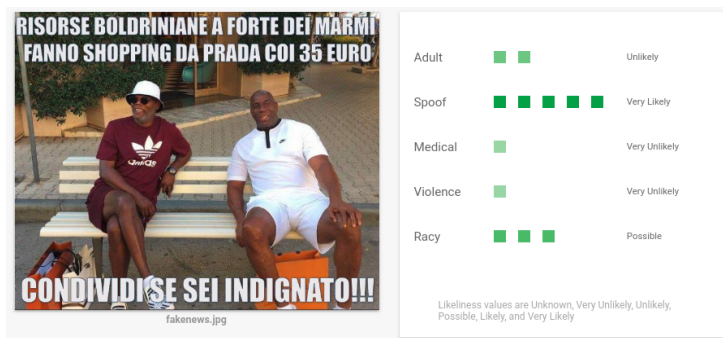


Figure 4.13: GCV: Safe Search results

in the image, categorization labeling by Google, the pages and links where the image is found, and safe search analysis.

Image's distribution on the web (DW)

Google Image Search retrieves all the links (it could be a web page or image source) where the searched image is used or stored. Specifically, the API divides all these links in 3 categories of images sources: Fully equal, Partially equal, and Web Pages where those fully/partially equal images are used. The features selected in this case are the quantities of the above-mentioned categories.

Safe Search (SS)

The SafeSearch analysis gives four evaluations according to the content of an image: Adult, Violence, Medical and Spoof. For each of them, a score that ranges from 1 to 5 is assigned. It is to be emphasized that the Spoof evaluation is profoundly based on the likelihood of the image being modified. An example of the GCV response for Safe search is reported below.

```
{
  "responses": [
    "safeSearchAnnotation": {
      "adult": "UNLIKELY",
      "spoof": "VERY_LIKELY",
      "medical": "VERY_UNLIKELY",
      "violence": "VERY_UNLIKELY",
      "racy": "POSSIBLE"
    },
  ]
}
```

Color Compositions (CF)

The API gives the 10 most dominant colors of an image, providing for each of them 2 types of value: score and pixel fraction. Pixel fraction, just as the name suggests, is the fraction of pixels that the color occupies in the analyzed image, while the score value is based on how much visual impact the color has, not taking into account how

much space it occupies. The processing selected is the calculation of the weighted mean of the R, G and B values of the dominant colors, having their relative score and pixel fraction as weights. At the end of this process, these are the 6 features to be included in the dataset: Impacts and Pixel Fractions of Red, Green and Blue in each image. A snippet of the GCV response for Color Composition service request is reported below.

```
{
  "responses": [
    {
      "imagePropertiesAnnotation": {
        "dominantColors": {
          "colors": [
            {
              "color": {
                "red": 78,
                "green": 37,
                "blue": 32
              },
              "score": 0.1583977,
              "pixelFraction": 0.046870183
            },
            {
              "color": {
                "red": 216,
                "green": 224,
                "blue": 249
              },
              "score": 0.038643196,
              "pixelFraction": 0.010792476
            },
            [...]
          ]
        }
      },
      [...]
    }
  ]
}
```

Faces and Emotions (FA)

Face recognition is one of the first interesting application of machine learning on images [88]. IN GCV there are 7 kinds of analysis for each detected face: joy, anger, sorrow, surprised, headwear, blurred, and under exposed. Each of them is given a degree of likelihood that varies from 0 to 6. In order to get more significant information from these features, aside from calculating the average degree for each emotion, the number of occurrences of strong emotions (with at least a 4 degree likelihood) is also taken into consideration. An example is reported below.

```
{
  "joyLikelihood": "VERY_LIKELY",
  "sorrowLikelihood": "VERY_UNLIKELY",
  "angerLikelihood": "VERY_UNLIKELY",
  "surpriseLikelihood": "VERY_UNLIKELY",
  "underExposedLikelihood": "VERY_UNLIKELY",
  "blurredLikelihood": "VERY_UNLIKELY",
  "headwearLikelihood": "POSSIBLE"
}
```

Detected Objects (DO)

This feature, just as its name suggests, tries to analyze what are the objects that can be found inside an image, and a score varying from 0 to 1 is assigned to every object detected. Obviously the images do not contain the same objects, hence, in order to have a well-aligned dataset, all the detected objects from all the images are collected and chosen as features: assigning its relative score if it is present in a given image, otherwise 0. An example of score obtained posting the previous image follows:

```
{
  "labelAnnotations": [
    {
      "mid": "/m/0b75wg4",
      "description": "photo caption",
      "score": 0.7388657,
      "topicality": 0.7388657
    },
    {
      "mid": "/m/04g3r",
      "description": "leisure",

```

```
    "score": 0.64685774,  
    "topicality": 0.64685774  
  },  
  {  
    "mid": "/m/06bm2",  
    "description": "recreation",  
    "score": 0.62521744,  
    "topicality": 0.62521744  
  },  
  [...]  
]
```

Web Entity tag (WE)

Similar to the previous one, this feature goes deeper in details; it tries to get the identity or even the characteristics of the object or person found in the image; it can even try to extract its main source (news source in this case). Of course all this information on Web entities is completely dependent on the knowledge of Google across the Web, hence, this might be considered as a biased feature. A web entity tag has also got its score, which can vary from 0 to an even larger number than 1. The procedure applied for aligning scores in object detection was also applied for this feature.

OCR

This operation has also been applied to the images in order to extract possible texts from them. Aside from the presence of text, a simple sentiment analysis has also been applied to the extracted texts, which gives two types of scoring: the negativity or positivity of the text that goes from -1 to 1, and the magnitude of the said sentiment that goes from 0 to any positive number. In our set of images, only 312 (177 hoax, 135 non hoax) have readable texts, which are relatively small numbers and also almost equally partitioned between the 2 classes. In fact, in the validation phase, the presence of OCR features would not affect the quality of the models. Of course, these data could still be used to fine tune the part of the system that takes care of different kinds of text analysis.

All the above features can be used in different kinds of classification models, in particular: Gaussian and Bernoulli Naive Bayes, Logistic Regression, Decision Trees, Random Forest and SVM. Classifications have also been applied to different subsets of the above mentioned features in order to investigate which subset has better performances.

Google Cloud Vision - API

Let us briefly introduce how to connect to GCV using the provided API. After installing the python Google Cloud client library and importing the correct module, the ImageAnnotator-Client class instances a client for sending requests to the corresponding Google service.

As preliminary activity, it is mandatory to register to the Google Cloud service in order to obtain the access token and to get the authorization for using the Google Cloud Platform.

An example of Google credential json file is the following:

```
{
  "type": "service_account",
  "project_id": "user-175914",
  "private_key_id": "d1e19cb541186688e623ad7de746d9a9b6794ad6",
  "private_key": "-----BEGIN PRIVATE KEY-----
MIIEvQIBADANBgkqhkiG9ASCBKcwg...ggSjAgEAAoIBAQDIHnZY
....
axwt5uxlBarvmAgE0Iz0nKX....N5eXZUYUfcEkgDdF5fIHved8
-----END PRIVATE KEY-----\n",
  "client_email": "prj@175914.iam.gserviceaccount.com",
  "client_id": "110480542773438953892",
  "auth_uri": "https://accounts.google.com/o/oauth2/auth",
  "token_uri": "https://oauth2.googleapis.com/token",
  "auth_provider_x509_cert_url":
    "https://www.googleapis.com/oauth2/v1/certs",
  "client_x509_cert_url":
    "https://www.googleapis.com/robot/v1/metadata/x509/
175914.iam.gserviceaccount.com"
}
```

This file must be exported as environment variables before sending queries, otherwise the service will be not accessible. This operation can be simply performed in a Linux based system with the following shell command

```
$ export GOOGLE_APPLICATION_CREDENTIALS="credential.json"
```

where the *credential.json* file contains the above mentioned data.

After these commands, it is possible to send requests to GCV with a few lines of code (see the next python snippet).

```
import io
import os
# Imports the Google Cloud client library
from google.cloud import vision
```



```
from google.cloud.vision import types

client = vision.ImageAnnotatorClient()

# The name of the image file to annotate
file_name = os.path.join(
    os.path.dirname(__file__),
    'images', 'fakenews.jpg')

# Loads the image into memory
with io.open(file_name, 'rb') as image_file:
    content = image_file.read()
image = types.Image(content=content)

# Performs safe search detection on the image file
response = client.safe_search_detection(image=image)

print(response)
```

The corresponding output is the following

```
safe_search_annotation {
  adult: UNLIKELY
  spoof: VERY_LIKELY
  medical: VERY_UNLIKELY
  violence: VERY_UNLIKELY
  racy: POSSIBLE
}
```

4.3.3 Image manipulation recognition

Another important approach that can be useful in fake information detection is the analysis of image manipulation. More precisely, it is very important to understand if a published image has been modified or not before being posted. Also in this field the CNNs are heavily used. In [7] the authors demonstrate that CNN can detect many editing operations on images, reaching a very high value in accuracy (i.e. 99.97%).

Figure 4.14 shows one of the 69 most famous viral fake news published in 2016.

Being able to annotate an image as "manipulated" when it is public on a social network can help to identify a possible fake news. We used this principle in our research in chapter 6.

Another very well known fake image is the diving shark during the hurricane Sandy (figure 4.15). The great matter is that, if not knowing the original image, one can understand



Figure 4.14: A famous example of fake news based on image manipulation



Figure 4.15: Another famous fake image: the Sandy hurricane shark

the manipulation only with particular techniques, used by researchers or by graphic designers [38].

Furthermore, in this case not even using GCV and the huge Google knowledge, it is possible to retrieve information related to an image authenticity. Posting the shark image to GCV service, just relating to Safe Search annotation, we obtain the following information

```
safe_search_annotation {  
  adult: VERY_UNLIKELY  
  spoof: VERY_UNLIKELY  
  medical: VERY_UNLIKELY  
  violence: UNLIKELY  
}
```

A normal Internet user could consider trusty the image, in a case like this. For this reason it is necessary to investigate also the relations that an image can have with other network elements (users, posts, websites) in order to increase the information about its authenticity. This aim can be reached by performing some network analysis.

4.4 Techniques for network analysis

This section illustrates the main analysis which can be applied to a social network. As reported in Sec. 2.2, a social network is usually represented as a graph where nodes symbolize persons and edges symbolize relations (i.e. there is an edge between node A and node B if A and B are connected by means of, e.g., friendship relation).

Formally we can define a social network as

$$G = (V, E)$$

where

V is the set of nodes (or *vertices*),

$$V = \{n_0, n_1, n_2, \dots, n_m\}$$

and E is the set of the edges

$$E = \{(xy) \mid x \in V \text{ and } y \in V \text{ and } \{there \text{ is a kind of relation between } x \text{ and } y\}\}$$

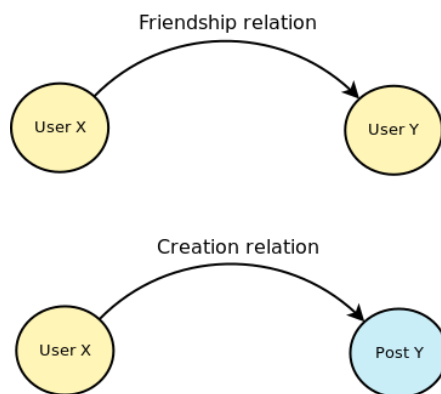


Figure 4.16: Different kinds of relations between nodes

In this work, nodes can represent not only persons but also all kinds of elements that occur in a social network, e.g. posts, websites, webpages, images, *likes* or *dislikes*, and, for this reason, a relation can have different semantic value if related to two users or, for example, to a user and a post.

In the first case, if X and Y were two users, the element (xy) would represent a *friendship* relation, if X were a user and Y a post, the element (xy) would represent a *creation* relation (i.e. the user X wrote the post Y).

In chapter 7 it will be explained all kinds of nodes and relations.

Now it will be introduced two of the most used operation which can be applied to a graph. In our study these have been used to retrieve information about the importance of single nodes,

4.4.1 Community detection

As highlighted in subsection 1.3.1, one of the most important operations to understand the existing dynamics between users in a social network is to analyze if, inside a larger community, strongly connected subgroups of users are present or not.

This operation is commonly called *Community Detection*. The most known algorithms for this purpose are

- Minimum-cut method
- Hierarchical clustering [54]

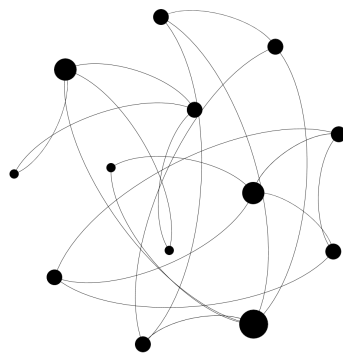


Figure 4.17: Social network graph



Figure 4.18: Communities detected in the same social network graph

- Modularity maximization
- Local-First method [3]

The aim of every community detection algorithm is to group the graph nodes in order to put together all those elements with similar characteristics or mainly connected: it depends by the kind of algorithm performed.

Fig. 4.17 and 4.18 show two graphs representing a Facebook user social network before and after applying the *Modularity Maximization* algorithm.

Modularity

A modularity algorithm for detecting community based on the relations between nodes was applied to the images and the result was that the three groups of nodes more connected each other have been grouped in the same community.

The value to be optimized is *modularity*, which is defined as a value in the range [-1, 1]. It measures the density of links *inside* communities compared to links *between* communities.

For a weighted graph, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where

A_{ij} represents the edge weight between nodes i and j ;

k_i and k_j are the sum of the weights of the edges attached to nodes i and j ;

$2m$ is the sum of all edge weights in the graph;

c_i and c_j are the communities of the nodes; and

δ is a simple delta function.

4.5 Conclusions

This chapter gave some details on the several machine learning techniques which have been used in this research, to find information about a news credibility. The discussed techniques are very general and are widely used for many different purposes. The next chapters of this report will focus only on the main problem of this work. All algorithms will be applied therein, in different contexts, to extract data related to trustiness of sources, posts, images, and all other aspects deemed useful for the task.

Part III

Credibility Discovery in Social Networks

Credibility evaluation system

This part of the dissertation contains a detailed description of the proposed system. The chapter starts with a brief introduction of the designed complete system, then it will introduce the different system components and their respective aims, finally it will explain what is new in our approach if compared to the existing literature.

A complex system for false information detection. Why?

The first question which may come in the reader's mind is: "Why do we need a complex system to identify fake news?". The answer could be that almost all piece of information, that is spread across the Web or social networks, consists of different aspects. Very rarely an information contains only text or only images: most often it contains both text and images and it is also related to sources like Web sites or social network elements or users.

For this reason it seems absolutely necessary to investigate all parts of an information to have a realistic estimation of its reliability. This approach has not been investigated in a systematic way in literature.

The complete system

Figure 4.19 describes the whole workflow of the proposed project: a post, but may be a different piece of information, is split in the four components (if they exist). For each component, a suitable analysis process is applied. This process could be a binary fake/no-fake classification process or a simple algorithm to retrieve a credibility estimation. In all cases the activated processes give out a trustiness value. All these outputs can be further combined to obtain a unique value of reliability for the information element.

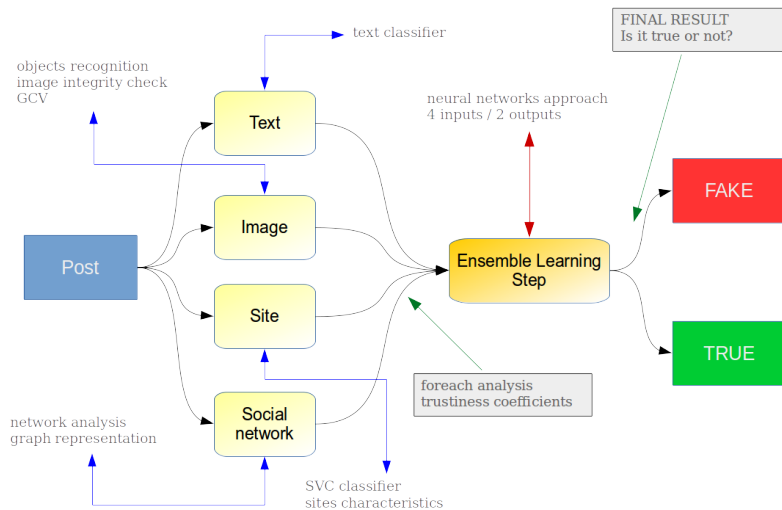


Figure 4.19: The complete process pipeline

Text analysis

For text analysis (see chapter 5) some classification algorithms have been tried. In this case the output is always a couple (*assigned class, confidence value*) where the assigned class can be "*fake*" or "*no-fake*" and the confidence value belongs to $(0, 1)$ range. Both these results can be used as inputs in the final decision step.

Image analysis

For image analysis (ref. chapter 6), each image has been analyzed looking for many characteristics, some intrinsic (colors, presence of objects...) and some extrinsic (i.e. related to external elements like how many times an image has been used in the web). This operation has been performed thanks to the Google Cloud Vision platform which permits to retrieve all this kind of information for each submitted image. After this step many experiments have been performed to identify the most significant features of fake detection system to be implemented.

Social network analysis

The third component of the entire process focuses on how an information is connected and linked to other elements of the Web, e.g., Web sites, Facebook pages, or mentioned users. In Chapter 7 a very deep analysis is made on how users, posts, pages and other social network elements are connected and how the study of these relations may be used for understanding the credibility of one of these elements. An iterative algorithm has been devised, to estimate a trustiness value working on a graph which represents relations between all social network items. Also, some classification and regression experiments have been performed, using data related to users' behaviors (see 7.1.4).

Sources structure analysis

The last part of the work shows how to retrieve information about a source trustiness (e.g., a Web site or a personal blog) just watching at some structural characteristics (e.g., presence of advertisements, number of links to other sites). Taking inspiration from the Stanford Guidelines for Web Credibility, many Web sites have been inspected looking for many aspects of these home pages and, after some preprocessing steps, each site has been represented as a set of interesting features used for training a classification system. Also in this case, like for text analysis, a confidence value in fake/no-fake class attribution can be obtained, with a significant precision level.

In the next four chapters, each of the above mentioned analysis is explained with details, from dataset collecting steps, to preprocessing operations, performed experiments, obtained results and comments. The last chapter shows some ideas for improving all the process and draws future perspectives.

Chapter 5

Text based trustiness analysis

Luckily, we still talk!

Even if the world is now tremendously pervaded by technology and human communication is more and more full of images, likes, emoticons, tags and many other contour elements (which are often unuseful), fortunately we still use our language to communicate. For this reason, text analysis is still one of the most important ways to understand what a human wants to tell to another human.

This chapter explains the approach taken in this work to this kind of analysis, related to trustiness evaluation based on texts spreaded in social networks. Following the text-analysis research principles shown in 4.1, this chapter explains the details of the process. First, it explains how data is collected from different sources and which data is used for this part of the research. The pre-process is also clarified, then the performed experiments are presented, and finally the results of text-analysis for credibility evaluations are shown and discussed.

5.1 Data collecting and preprocessing

Typically, the first step of a research is to find relevant data. In this case, the problem was the opposite: there are too many textual data on the Web which would be analyzed to discover information. In particular, this work has focused on social networks and Web sites, which are still the most used “virtual places” where users search information, especially for text.

In this part of the research, a dataset has been built downloading many tweets by Twitter ¹ accounts. This choice has been dictated by the fact that several researches use Twitter data for

¹Twitter inc. <https://www.twitter.com>

this purpose and so it is possible to find many free datasets on the Web. The chosen approach in this case has been to build one self-made dataset and to download a public one, for the sake of comparison.

5.1.1 Public dataset

The public dataset (DS1 in the following), downloaded from the FakeNewsNet channel ² [78], is composed by 210 real json-formatted tweets and 210 fake ones, both speaking about political facts.

From each tweet, only text elements have been used for our experiments. They have been pre-processed as follows:

Text preprocess

Let us illustrate how pre-processing works, starting from the following example:

```
16.8k SHARES SHARE THIS STORY
```

```
Hillary Clinton just called out the fact that Donald Trump cheered for the housing crisis in anticipation of its collapse \u2013 which is absolutely true. Trump told The Globe and Mail in March of 2007: \u201cPeople have been talking about the end of the cycle for 12 years, and I\u2019m excited if it is. I\u2019ve always made more money in bad markets than in good markets.\u201d
```

```
In fact, Trump thought the housing crisis was much-ado-do about nothing for high-end investments, and told investors: \u201cI don\u2019t see the subprime problems affecting the higher-end stuff\u2026In fact, he is advising investors that there are now great deals in buying subprime mortgages at a discount and repossessed houses at low prices.\u201d
```

```
Of course, on one level, Trump wasn\u2019t wrong\u2014\u2014\u2014in that the subprime crisis affected mostly poorer Americans who lost their homes and jobs and the ability to support their families. The collapse of home prices caused by the housing bubble cost roughly seven million Americans more than $7 trillion in equity during the Great Recession.
```

```
This recession most severely impacted low-income folks \u2013 people who are burdened with payments in excess of 50 percent of their income \u2013 Trump\u2019s response is just \u201cThat\u2019s called business.\u201d
```

²<https://github.com/KaiDMML/FakeNewsNet>


Emoji	Unicode	Description	Token
	U+1F603	smiling face with open mouth	SMILE

Figure 5.1: Text representation of an emoticon

Add your name to millions demanding that Congress take action on the President's crimes. IMPEACH DONALD TRUMP!

Pre-processing encompasses three steps: (i) *stop-words filter*, aimed at removing non significant words like pronouns, articles, conjunctions and similar; (ii) stemming and lemmatization (see 4.2.1); (iii) ad-hoc items removal for all Twitter specific elements like entities, mentions, urls and substitution of emoticons with their corresponding meaning (see fig. 5.1).

After applying these operations, the given text has been turned into the following one:

16 8k share share thi stori hillari clinton just fact donald trump cheer
 hous crisi anticip collaps absolut true trump told globe mail march 2007
 end cycl 12 years, i'm excit i'v alway money bad market good market
 fact, trump thought hous crisi wa much-ado-do noth high-end investments
 told investors don't subprim problem affect higher-end stuff fact
 advis investor great deal buy subprim mortgag discount reposess hous
 low price Of course, level, trump wasn't wrong subprim crisi affect
 mostli poorer american lost home job abil support famili collaps home
 price caus hous bubbl cost roughli seven million american \$7 trillion
 equiti dure great recess thi recess sever impact low-incom folk peopl
 burden payment excess 50 percent incom trump respons just busi add
 million demand congress action president' crime impeach donald trump

Fig. 5.2 reports a small part of the public dataset, after preprocessing.

5.1.2 Self-made dataset

To compare results, a complete new dataset has been built (DS2 in the following), downloading tweets from Twitter. To obtain a well-balanced dataset, downloaded data include 527 tweets published by accounts which were indicated as not trusty from Twitter community (almost all of them have been now removed or blocked by the social network administrator) and 1928 tweets posted by many credible accounts as newspapers, journalists, tv-broadcaster. The former are considered “fake”, and the latter “no-fake” (true).

	text	class
86	peopl notic someth odd about hillari outfit at...	FAKE
87	olymp committe buckl to lgbt groups... make ga...	FAKE
88	it was late one night in the white hous when o...	FAKE
89	oh hillari when will you learn? you never coun...	FAKE
90	5.2k share facebook twitterjason falcon the nr...	FAKE
91	hillari clinton campaign is make one plea ahea...	TRUE
92	share share facebooktwitt googlepinterestdiggl...	TRUE
93	unit nation cnn presid barack obama made an im...	TRUE
94	donald your a snivel coward ted cruz said in m...	TRUE
95	as polic today captur the man want for questio...	TRUE

Figure 5.2: Example of self-built textual dataset

The dataset has been balanced using only 500 elements per class, thus obtaining a corpus of 1000 total tweets.

As in the previous case, the same preprocessing steps have been applied to our dataset.

5.2 Data analysis and classification

Several classification algorithms have been applied on the cited datasets, after splitting each one in train- and test-set. Both datasets have been divided using 66% of samples for training and the remaining ones for testing.

Vectorization and bag-of-words model

Datasets have been processed to build the bag-of-words model.

DS1 has been transformed in a dataframe with 16650 columns, each representing the presence or absence of a word inside a tweet. DS2 instead has only 5732 columns, meaning that the underlying vocabulary is shorter.

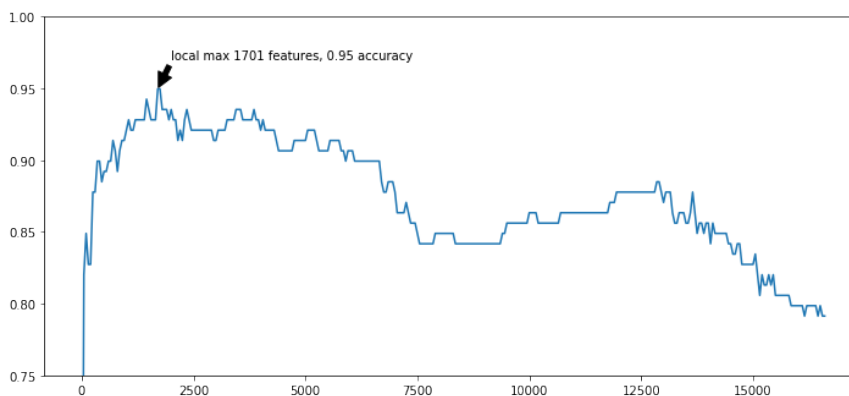


Figure 5.3: Accuracy according different features number for DS1

Features selection

Not all the features have the same importance in classification. This can be discovered applying the *tf/idf* function to estimate the *weight* of each feature in discriminating the different classes. For this reason we have calculated the accuracy using a different number of features. Fig. 5.3 shows that the best result for DS1 in terms of accuracy (95%) is obtained using only 1701 features. For DS2, instead, the best performance (89%) has been calculated using 670 features.

5.3 Results and conclusions

Considering the previous result, the comparison between different classifiers has been estimated using only the best subset of features.

Algorithm comparison

Table 5.1 shows the results obtained applying different algorithms. As often found in the scientific literature, the Random Forest and Naive Bayes Multinomial classifiers have provided good results in terms of classification accuracy.

Considering that Naive Bayes Multinomial classifier is the fastest in terms of training time, and that results are not so different from the other classifiers, as in text-analysis literature, this algorithm has been selected for this part of the project.

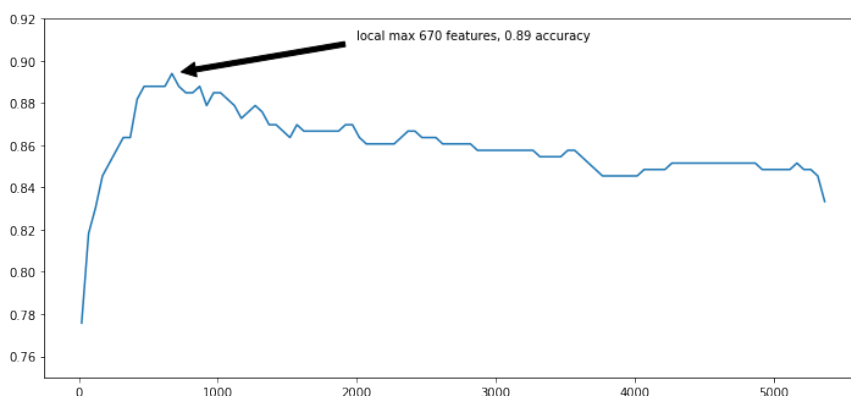


Figure 5.4: Accuracy according different features number for DS2

Table 5.1: Comparison of algorithms in the classification accuracy for self-build dataset

Algorithm	Accuracy	Precision	F-score
MLP - Neural Network	86.6%	85.3%	86.9%
Random Forest	83.3%	85.2%	82.8%
SVM	83.3%	92.3%	81.3%
KNN	50.3%	50.1%	66.8%
Naive Bayes	81.8%	94.8%	78.7%

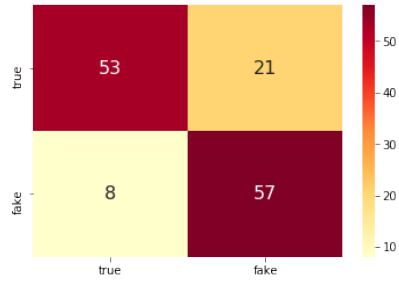


Figure 5.5: Confusion matrix calculated on FakeNewsNet dataset DS1

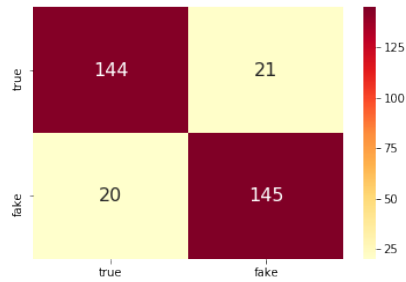


Figure 5.6: Confusion matrix calculated on dataset DS2

Fig. 5.6 and 5.5 show the confusion matrices related respectively to DS1 and DS2 dataset, respectively.

Chapter 6

Image based trustiness analysis

WYSIWYG. It's not true on the Web!

This chapter is focused on the images attached and posted on the Web, in particular in social networks.

The spreading of simple technologies for image manipulation, combined with the easiness in information dissemination, has rapidly generated a great problem for assessing the credibility level of news sources.

This part of the work investigates in-depth the structure of an image and has the aim to find the common characteristics (if any) of images used in fake news.

This dissertation does not deal with the detection of manipulated images, because in several cases a false information is spread with a true image (i.e. no changes are made on it before its usage). For example, for disseminating a fake information about a certain politician, often it is associated with images where he/she has a ugly face, a bad look or ridiculous body positions. It means that in many cases the image is real, but it is used in a distorted way.

Nevertheless, the experiments have shown that it is possible to identify a false information with a fair precision just analyzing the structure of linked images.

6.1 Data collecting and preprocessing

The first problem faced for this kind of analysis is to create a useful data set. The research proceeded in the same way that has been discussed in the previous chapter, starting from a

list of Facebook pages, half reported as unreliable by different independent sources ¹ and half considered reliable, as science or magazines pages ².

More than 2000 posts containing images have been downloaded, about 1000 from untrusted pages and others from trusted ones.

Feature extraction

As already explained in Subsection 4.3.2, all images have been submitted to GCV system, retrieving for each image all the information concerning

- objects occurring inside (DO)
- color composition (CF)
- presence or absence of faces and eventually the face emotions (FA)
- texts that may occur inside (OCR)
- safe search tags (SS)
- Web entities (WE)
- distribution on the Web (DW)

As reported in the cited subsection, the first four features are totally objective and do not depend on Google Knowledge Base (GKB). The last three, instead, are absolutely dependent on the Google Knowledge Base and have been used, in our experiments, to compare our results with Google fake information detection tool.

Classification experiments and results

Different classification experiments have been performed, with different algorithms and different features sets.

Tables 6.1, 6.2, 6.3 and 6.4 report different results.

¹<http://www.butac.it/> for Italian language pages and <https://www.snopes.com/> for English language pages.

²E.g. <https://www.facebook.com/cnn/> or <https://www.facebook.com/ilsole24ore/> respectively for English and for Italian languages.

Table 6.1: Classification results using ALL features

Model	All Features		
	Precision	Recall	F1-Score
Gaussian	0.74	0.81	0.77
Bernoulli NB	0.83	0.83	0.83
Logistic Regression	0.78	0.62	0.69
Decision Trees	0.71	0.71	0.71
Random Forest	0.77	0.86	0.81
SVM	0.54	0.62	0.57

Table 6.2: Classification results using only web entities

Model	Web Entities (WE)		
	Precision	Recall	F1-Score
Gaussian	0.77	0.82	0.79
Bernoulli NB	0.82	0.87	0.84
Logistic Regression	0.88	0.58	0.70
Decision Trees	0.80	0.69	0.74
Random Forest	0.81	0.92	0.86
SVM	0.77	0.65	0.71

Table 6.3: Classification results using only detected objects

Model	Detected Objects(DO)		
	Precision	Recall	F1-Score
Gaussian	0.51	0.75	0.61
Bernoulli NB	0.67	0.68	0.68
Logistic Regression	0.66	0.63	0.65
Decision Trees	0.59	0.66	0.62
Random Forest	0.66	0.73	0.70
SVM	0.73	0.56	0.64

Results comment

Results reported in Tables 6.1, 6.2, 6.3 and 6.4 allow to make some observations: first, the great impact that Google Knowledge Base information has on classification results. In terms of accuracy, if just a group of feature is considered, the best performances are in fact reached

Table 6.4: Classification results using Color factors

Model	Color Factors (CF)		
	Precision	Recall	F1-Score
Gaussian	0.51	0.69	0.58
Bernoulli NB	0.48	0.99	0.65
Logistic Regression	0.54	0.64	0.59
Decision Trees	0.49	0.52	0.51
Random Forest	0.48	0.53	0.50
SVM	0.50	0.67	0.57

using Web Entities. It means that GKB information is extracted by a best trained classification system which, most likely, uses many other techniques for assigning a credibility evaluation to a certain image ³.

However, for evaluating the system described here, it must be independent by an external knowledge base in order to be absolutely based only on objective data related to the analyzed image.

Anyway, if one considers, for a complete system, the results obtained using only the Detection Object feature group and Random Forest algorithm, 70% accuracy in classification can be reached.

This result is very interesting, because it means that, only looking to image composition (i.e. the objects embedded inside an image) it's possible to have a perception of the credibility of the related information. This is probably due to the fact that the authors of fake news usually use the same kind of images, for example images with scary faces, blood, screaming persons, crowd, coloured people, malnourished children and many other strong elements to hit the reader and to convince him to believe and share the news.

³For example Google Claim review system allows users to submit data for fact-checking.

Chapter 7

Social graph based trustiness analysis

“A man is known by the company he keeps” (Aesop)

Aesop’s citation tells us a great truth. Since ancient times the personal reputation of a man has been evaluated also by looking at the people closest to him; and this approach has not been affected by modern technologies. In fact, it has become even more evident. Many researches, e.g. [81], show that inspecting a user’s relations in social network can give many information about his/her credibility and his/her behaviors. This chapter explains a methodology that, starting from the analysis of many trusty Facebook pages and many unreliable pages, allows to estimate, with a quite good precision, the credibility level of a social user.

7.1 Collecting data about social network users

As usual, the first step has been focused on retrieving information about users and published posts, in particular on existing relations among them. Differently from other parts of the project, Social Network Analysis has not been performed just using classification algorithms; rather the dataset has been built looking for relations between users, posts, pages and external sources, when present, connecting each other.

During data retrieval, some of the following questions have been considered:

- Who created a post?
- Which users are related to the news creator?

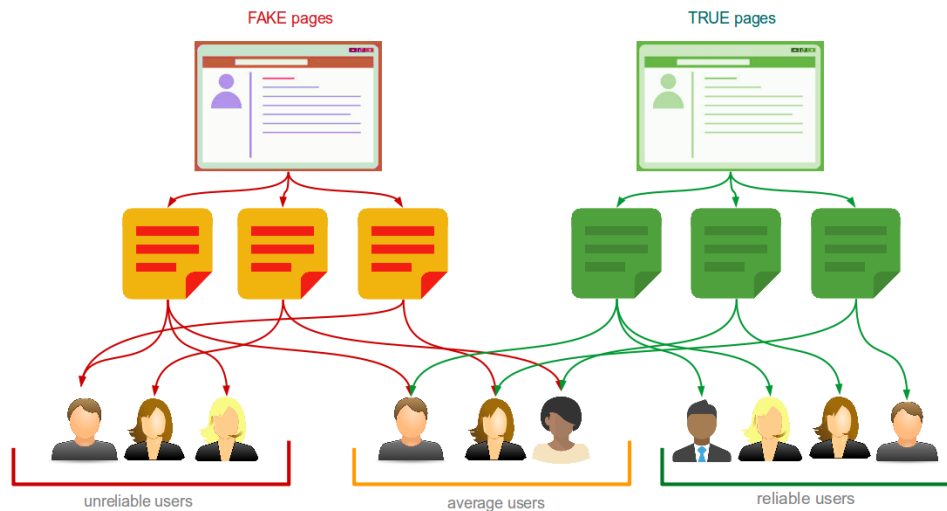


Figure 7.1: different users groups descrimination

- Which are the pages that the user follows?
- What activity level does the user show on the social network?
- Which posts does the user share?

Answering these questions allows to inspect a user's behavior and, consequently, to assign him/her a credibility value. The next paragraphs will explain the process followed for this aim.

7.1.1 Retrieving trusted and untrusted users

Referring to Section 6.1, at first 50 Facebook pages have been analyzed, equally divided between reliable and unreliable accounts. From each page, the last 1000 posts have been analyzed to look for users who shared them.

In cases in which posts contain links to Web sites, or other page references, the corresponding information has been stored in a database.

After performing this operation, this data have been collected: 47274 posts, 236 new Facebook pages, 530 Web sites and 300384 Facebook users.

This information has been represented with a graph (see 7.2) with 348897 nodes and 752711 edges.

As nodes represent different concepts (e.g. users, posts, sites, pages), also edges have different meanings, according to the values of the nodes they connect.

The following edges have been used:

- *sharings*: connecting a user-node and a post-node;
- *publications*: connecting a page-node and a post-node;
- *citations*: connecting a post-node and a site-node or two page-nodes;
- *mentions*: connecting two user-nodes.

7.1.2 Initial credibility value

After downloading data, a default credibility value has been assigned to each elements directly retrieved from the initial 50 pages.

- **0.1** for all posts published by a not reliable page
- **0.9** for all posts published by a trusted page
- **0.1** for those users (called "unreliable") which shared only posts from fake pages
- **0.5** for those users (called "average") which shared posts from both the categories
- **0.9** for those users (called "reliable") which shared only posts from true pages

All other elements, not directly connected to initial pages, have been initialized with **-1** value;

7.1.3 Reliability estimation

In such a graph, nodes contain also the most important value of our research: the reliability value. We have developed an iterative algorithm (see Alg. 1), which calculates the reliability of all nodes, starting from few available values.



Figure 7.2: A graph representation of downloaded data

Algorithm 1: Iterative algorithm to estimate credibility of all elements in graph

Data: social graph

Result: social graph with all evaluated nodes

1 V is the array which contains the reliability value of all nodes;

// V initially contains only the already set values

// for some users, posts, pages

// The not calculated values are set to -1 by default

2 δ represents the V variation between two iterations;

3 ϵ is a minimal bound to stop iterations;

4 $G = (N, E)$ represents the graph;

5 r_i is node i credibility value ;

6 w_i is node i weight;

// $w=100$ for page-nodes and site-nodes

// $w=10$ for post-nodes

// $w=1$ for user-nodes

7 N_k represents all already valorized nodes connected to node k ;

8 **while** (there are still no evaluated nodes) $\parallel (\delta < \epsilon)$ **do**

9 $V' \leftarrow V$;

10 **foreach** $k \in N$ **do**

11 $V'[k] \leftarrow \sum_x \frac{x_i w_i}{w_i}$;

12 $\delta \leftarrow \|V - V'\|$

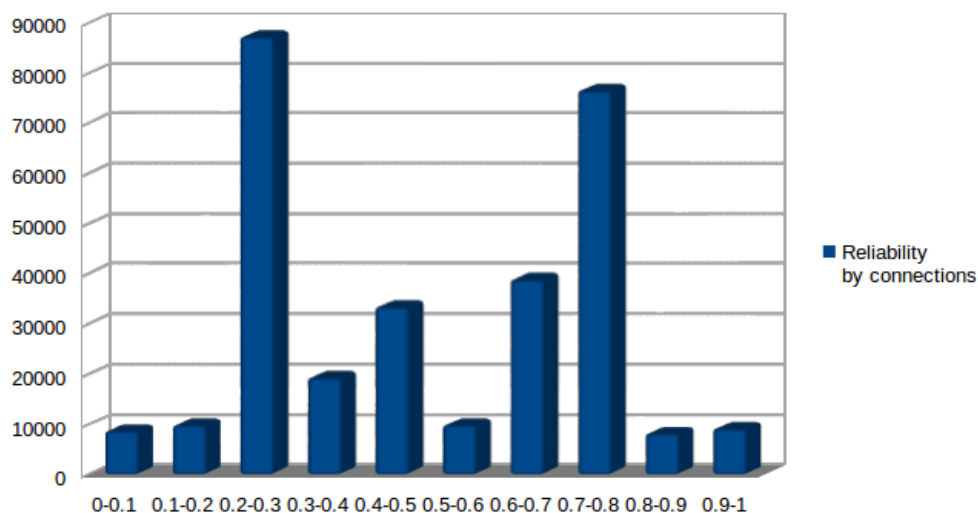


Figure 7.3: Reliability calculated by social graph connections

At the end of the algorithm, all nodes in graph G are valorized with a certain value of credibility.

Fig 7.3 shows the credibility level distribution among users. As expected, the distribution has two peaks around 0.25 and 0.75 values. It means that, just following which posts are shared, it's possible to divide users in two categories: trustworthy and not trustworthy. A similar chart (see Fig 7.4) shows the calculated values for all considered Facebook pages.

7.1.4 Reliability estimation by behavior

The second interesting result of this research has been that it's possible to assign a credibility value to users, looking only at the categories of pages he/she likes (i.e. Sport, Magazine, Politics, Society, Music and so on...), even without the knowledge of those precise pages.

For this experiment the 3000 most active users¹ have been considered (1000 for each of three above mentioned groups).

For each user in this group, the Facebook bulletin board has been analyzed (if visible and not private), downloading posts and shares. This has allowed to understand which were the categories of pages shared from each user, giving him/her a newly defined "*n-interaction*" value.

¹Most active users are those which have the highest number of interactions with pages

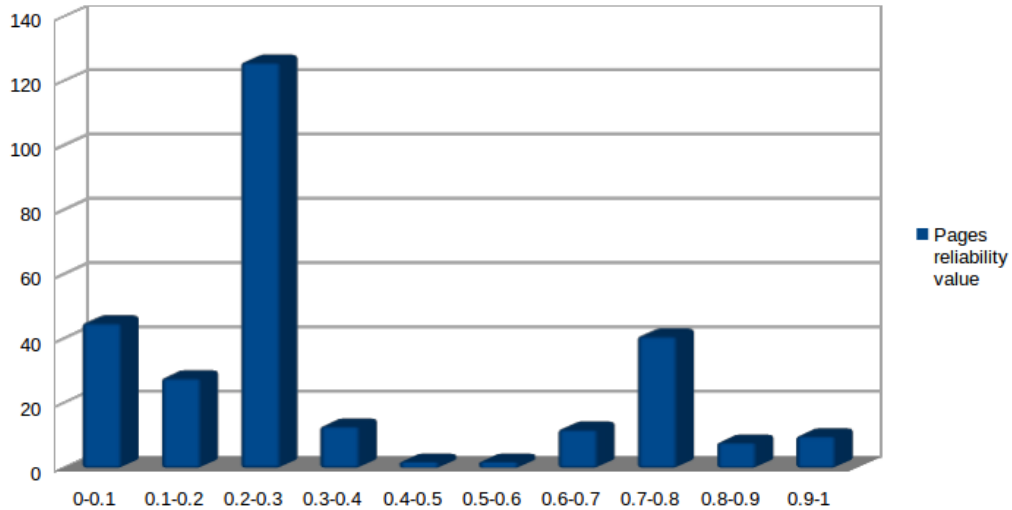


Figure 7.4: Reliability calculated for Facebook pages

In this way, for these 3000 users, it has been calculated how many interactions they have had with pages of a certain category (i.e. a user U has n -interaction with category C , if he/she has shared or liked n times posts published on pages which belong to category C).

This case of study can be correctly depicted by a bipartite graph, because neither users nor pages have any interactions with nodes of the same type. Fig. 7.5 shows a bipartite graph where edges are also labeled with the number I_{ij} of interaction between users i and page j .

The same information can also be represented as a table where rows denote users and columns categories (see Table 7.1).

7.1.5 Classification

At this point, the last two experiments on this kind of data have been performed: a classification test for understanding to what extent the user behavior (seen as which kind of pages a user follows) can be used to classify him/her as trusted or untrusted.

Considering a user as an array of 880 numeric features (e.g. $U1$ in Tab 7.1 is represented as $[3, 0, \dots, 6, 0]$ array), all dataset instances (3000, as users) have been divided between training and test set. Both are made up of 1500 instances, 500 for each existing users class (i.e. *unreliable*, *average* and *reliable*).

Different classification algorithms with different number of features (selected by ANOVA

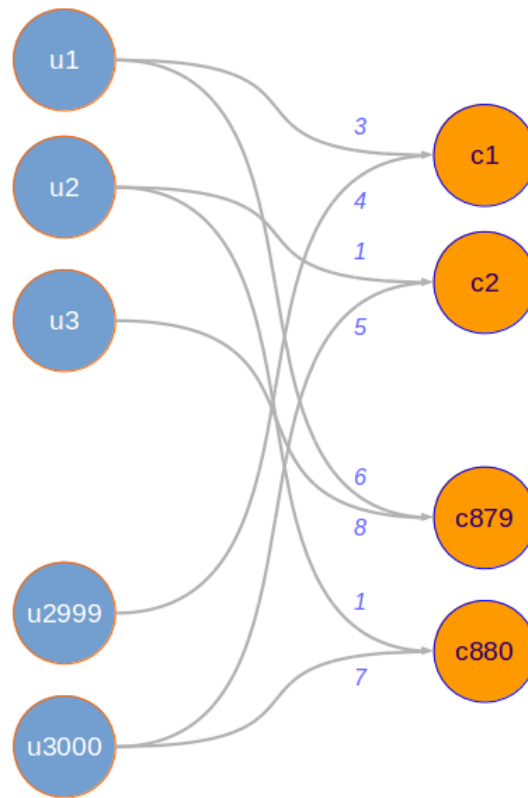


Figure 7.5: bipartite graph of users-pages interactions

Table 7.1: Interaction between users and categories

User	reliability	class	cat 1	cat 2	...	cat 879	cat 880
<i>U1</i>	0.32	fake	3	0	...	6	0
<i>U2</i>	0.27	fake	0	1	...	0	1
<i>U3</i>	0.69	true	0	0	...	8	0
...
<i>U2999</i>	0.79	true	4	0	...	0	0
<i>U3000</i>	0.41	fake	0	5	...	0	7

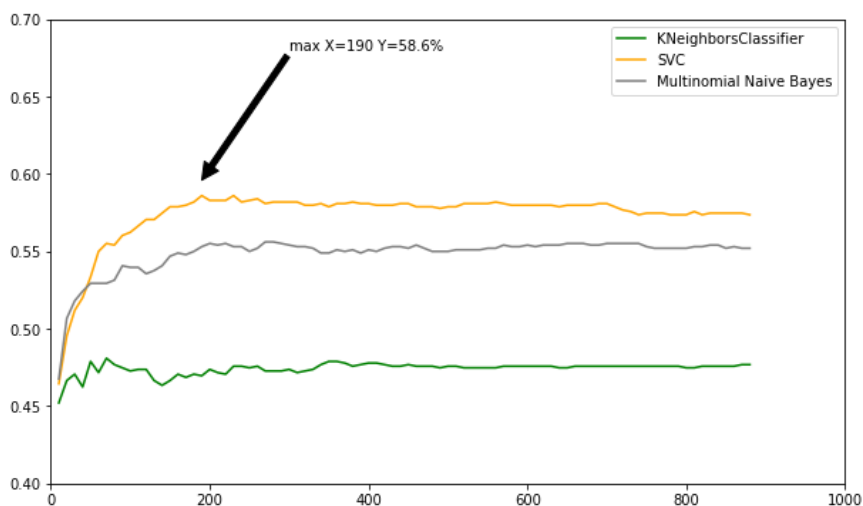


Figure 7.6: Comparison between algorithms in classification

algorithm) show that a quite good accuracy (almost 60% with also 190 analyzed categories) can be obtained just looking at a user's behavior. In Figure 7.6 a comparison between different classifiers is shown.

7.1.6 Credibility prediction by regression

The last experiment we have performed is to predict the credibility value with a linear regression model. The same above mentioned data set has been used for this test. The graph in Figure 7.7 shows that about half predicted values have quite the correct calculated value. It means that also a regression approach can be used for estimating the credibility of a user, looking just at the actions that he/she performs on the social network.

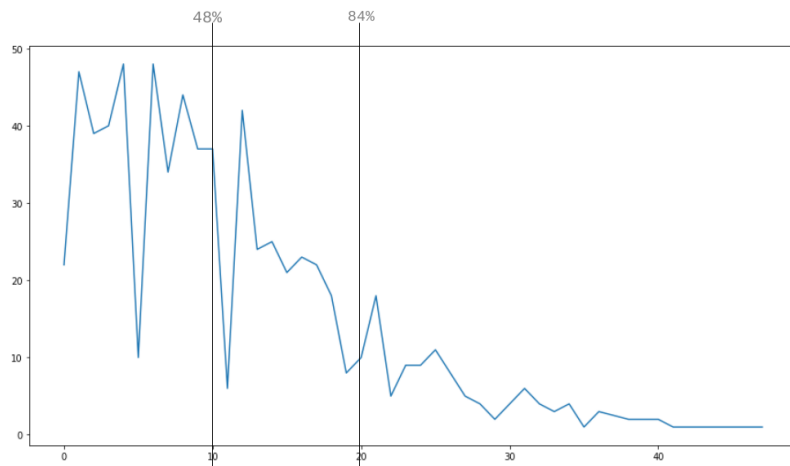


Figure 7.7: Regression estimation errors. About half predicted results differ less than 0.1 from the values calculated with Alg. 1

Chapter 8

Source structure based detection system

Appearances can be deceiving

This chapter explains the last part of the project, which focuses only on the analysis of site structures, to discover any useful information for the credibility problem. In fact, looking only at how Web sites are built and developed, it can be observed that fake news sites have similar characteristics, which can be used for assigning a reliability value also without analyzing the published contents.

In this part, the published texts or articles writing style have not been considered, because this information is related to text-analysis which has been detailed in Chapter 5. Rather, this part of the analysis looks for well-known properties which a good web site should always have.

We have started referring to the Stanford Guidelines for Web Credibility¹ which say that to “*boost a Web site’s credibility*” a developer should

1. Make it easy to verify the accuracy of the information on your site.
2. Show that there’s a real organization behind your site.
3. Highlight the expertise in your organization and in the content and services you provide.
4. Show that honest and trustworthy people stand behind your site.

¹<https://credibility.stanford.edu/guidelines/index.html>

5. Make it easy to contact you.
6. Design your site, so that it looks professional (or it is appropriate for your purpose).
7. Make your site easy to use – and useful.
8. Update your site's content often (at least show it's been reviewed recently).
9. Use restraint with any promotional content (e.g., ads, offers).
10. Avoid errors of all types, no matter how small they seem.

8.1 Collecting data about web sites

Paying attention to these rules, the structure of many Web sites has been analyzed, looking for those elements that are related to the guidelines.

Relying on the same list of sites identified for image data collecting (see Section 6.1), the home pages of 160 web sites has been downloaded – 85 labeled as reliable and 75 as unreliable.

Features extraction

The structure of Web sites is analyzed, looking for the following features, which are inspired to the above mentioned guidelines. The most significant features are:

- self-defined satirical site (boolean feature);
- kind of developing software (well-known cms software like wordpress or home-made software);
- number of donation links or button (integer value);
- number of occurring advertisements;
- presence of liability disclaimer (boolean feature);
- presence of VAT number;
- average number of daily posts (i.e., the update rate);
- number of links in home page;
- number of self-referred links;
- number of external links.

Figure 8.1 shows an example of collected data.

	satiric	software	donate_buttons	disclaimer	VAT	dailyPostsNumber	totalLinks	selfLinks	CLASS
17	0	0	0	0	0	34	295	3	T
18	0	0	1	0	0	3	977	143	T
19	0	0	0	0	0	214	311	126	T
20	0	0	0	0	0	14	527	173	T
21	1	1	1	0	0	0	79	10	F
22	0	1	0	0	0	0	1030	1022	F
23	0	1	0	0	0	0	25	21	F
24	1	0	0	0	0	1	110	85	F
25	1	0	0	0	0	0	1080	1073	F

Figure 8.1: Web sites features data set example

Table 8.1: Comparison of algorithms in the source analysis classification accuracy

Algorithm	Precision	Recall	F-score
MLP - Neural Network	0.79	0.79	0.79
Random Forest	0.83	0.83	0.83
SVM	0.76	0.70	0.68
KNN	0.82	0.79	0.78
Naive Bayes	0.77	0.70	0.68

Classification experiments and result

As reported in the previous chapters, after building the datasets, different classification experiments have been performed, using various algorithms. Table 8.1 shows that Support Vector Machine and Multilayer Perceptrons allow to obtain very good performances in classification, whereas Table 8.2 reports the 5 most significant features, together with their respective information value.

The last table (i.e., Table 8.3) shows the classification results obtained using only the 5 most significant features mentioned above.

Table 8.2: Most significant features in source structure based classification

feature	value
software	16.68
dailyPostsNumber	8.19
totalLinks	6.54
selfLinks	2.10
satiric	1.31

Table 8.3: Comparison of algorithms in the source analysis classification accuracy using only the features reported in Table 8.2

Algorithm	Precision	Recall	F-score
MLP	0.88	0.88	0.88
Random Forest	0.83	0.83	0.83
SVM	0.81	0.73	0.70
KNN	0.83	0.83	0.83
Naive Bayes	0.87	0.84	0.83

Conclusions

These results allow to assert that also the structure of sources mentioned in pieces of information deserves to be considered, in the task of fake news identification. Indeed, the results highlight that, just paying attention to few characteristics of the home page, it is possible to estimate the reliability of a source with a not negligible accuracy.

Future works

A complete trustworthiness automatic detection system

The functioning of the single parts of the project, for automatic fake news detection, has been explained in the previous chapters, in detail. Four different subsystems have been indeed shown, for figuring out the problem of false information detection, analyzing: *(i)* texts, *(ii)* images, *(iii)* social network relations, and *(iv)* cited sources structure. Each one of these subsystems can estimate a level of credibility for a different aspect of the studied piece of information.

The next step, which should be made during future work, is to put all these different classification outputs together, to build an integrated system capable of evaluating a given input instance (e.g. a Facebook status posted by a certain user, with an image and some cited sites) and emitting a numeric estimation of its trustiness.

Finally, the whole system should also provide a public API to allow external systems to interact with its knowledge base, both for making queries and for sending users' feedbacks.

In this way, the complete system could be also ready to learn continuously about new instances, improving its performances.

Bibliography

- [1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
- [2] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.
- [3] M. Amoretti, A. Ferrari, P. Fornacciari, M. Mordonini, F. Rosi, and M. Tomaiuolo. Local-first algorithms for community detection. In *KDWeb*, 2016.
- [4] G. Angiani, A. Ferrari, P. Fornacciari, M. Mordonini, and M. Tomaiuolo. Real marks analysis for predicting students’ performance. In *International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning*, pages 37–44. Springer, 2018.
- [5] G. Angiani, P. Fornacciari, E. Iotti, M. Mordonini, and M. Tomaiuolo. *Models of Participation in Social Networks*, pages 196–224. IGI Global, 2017.
- [6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
- [7] B. Bayar and M. C. Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11):2691–2706, 2018.
- [8] A. Bessi and E. Ferrara. Social bots distort the 2016 us presidential election online discussion. 2016.
- [9] A. Beutel, K. Murray, C. Faloutsos, and A. J. Smola. Cobafi: collaborative bayesian filtering. In *Proceedings of the 23rd international conference on World wide web*, pages 97–108. ACM, 2014.

- [10] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 119–130. ACM, 2013.
- [11] S. P. Borgatti and P. C. Foster. The network paradigm in organizational research: A review and typology. *Journal of management*, 29(6):991–1013, 2003.
- [12] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102, 2014.
- [13] R. S. Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology*, 92(6):1287–1335, 1987.
- [14] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Icwsn*, pages 61–70, 2015.
- [15] A. Clementi, P. Crescenzi, C. Doerr, P. Fraigniaud, F. Pasquale, and R. Silvestri. Rumor spreading in random evolving graphs. *Random Structures & Algorithms*, 48(2):290–312, 2016.
- [16] J. S. Coleman. Social capital in the creation of human capital. *American journal of sociology*, 94:S95–S120, 1988.
- [17] N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 82:1–82:4, Silver Springs, MD, USA, 2015. American Society for Information Science.
- [18] B. Doerr, M. Fouz, and T. Friedrich. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75, 2012.
- [19] J. S. Donath. Identity and deception in the virtual community. In *Communities in cyberspace*, pages 37–68. Routledge, 2002.
- [20] P. Dwyer. Measuring collective cognition in online collaboration venues. *International Journal of e-Collaboration (IJeC)*, 7(1):47–61, 2011.
- [21] P. Erdős and A. Rényi. On random networks. *Pub. Math*, 6:290–297, 1959.
- [22] D. Fallis. A functional analysis of disinformation. *iConference 2014 Proceedings*, 2014.
- [23] J. Fedorowicz, I. Laso-Ballesteros, and A. Padilla-Meléndez. Creativity, innovation, and e-collaboration. *International Journal of e-Collaboration (IJeC)*, 4(4):1–10, 2008.

- [24] U. Feige, D. Peleg, P. Raghavan, and E. Upfal. Randomized broadcast in networks. *Random Structures & Algorithms*, 1(4):447–460, 1990.
- [25] A. J. Flanagin and M. J. Metzger. Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, 77(3):515–540, 2000.
- [26] P. Fornacciari, M. Mordonini, and M. Tomaiuolo. A case-study for sentiment analysis on twitter. In *WOA*, pages 53–58, 2015.
- [27] P. Fornacciari, M. Mordonini, and M. Tomaiuolo. Social network and sentiment analysis on twitter: Towards a combined approach. In *KDWeb*, pages 53–64, 2015.
- [28] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
- [29] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *The International Journal of High Performance Computing Applications*, 15(3):200–222, 2001.
- [30] N. Fountoulakis, K. Panagiotou, and T. Sauerwald. Ultra-fast rumor spreading in social networks. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1642–1660. SIAM, 2012.
- [31] 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.
- [32] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.
- [33] J. Fulk, R. Heino, A. J. Flanagin, P. R. Monge, and F. Bar. A test of the individual action model for organizational information commons. *Organization Science*, 15(5):569–585, 2004.
- [34] 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.
- [35] Gazzetta. Codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e scientifici - deontological and good practice code for treating personal data used in scientific and statistical issues. *Gazzetta Ufficiale della Repubblica Italiana*, 2004(190), 2004.
- [36] J. Golbeck and J. Hendler. Inferring binary trust relationships in web-based social networks. *ACM Transactions on Internet Technology (TOIT)*, 6(4):497–529, 2006.

- [37] M. Granovetter. 1973: The strength of weak ties, *American Journal of Sociology* 78, 1360-1380. 1973.
- [38] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.
- [39] V. Gurusamy and S. Kannan. Preprocessing techniques for text mining, 10 2014.
- [40] C. Harris. Detecting deceptive opinion spam using human computation. In *Workshops at AAAI on Artificial Intelligence*, pages 87–93, 2012.
- [41] S. C. Hayne and C. Smith. The relationship between e-collaboration and cognition. *International Journal of e-Collaboration (IJeC)*, 1(3):17–34, 2005.
- [42] P. Herson. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995.
- [43] 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.
- [44] H. Hoang and B. Antoncic. Network-based research in entrepreneurship: A critical review. *Journal of business venturing*, 18(2):165–187, 2003.
- [45] P. W. Holland and S. Leinhardt. The statistical analysis of local structure in social networks, 1974.
- [46] B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, D. Makhija, and C. Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 495–503. SIAM, 2016.
- [47] P. N. Howard and B. Kollanyi. Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. 2016.
- [48] J. Jacobs. The uses of sidewalks: safety. *The City Reader*, pages 114–118, 1961.
- [49] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Inferring strange behavior from connectivity pattern in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 126–138. Springer, 2014.
- [50] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230. ACM, 2008.

-
- [51] A. Klein, H. Ahlf, and V. Sharma. Social activity and structural centrality in online social networks. *Telematics and Informatics*, 32(2):321–332, 2015.
- [52] S. Kumar and N. Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.
- [53] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.
- [54] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.
- [55] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1063–1072. International World Wide Web Conferences Steering Committee, 2017.
- [56] J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1566–1576, 2014.
- [57] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3367–3375, 2015.
- [58] G. C. Loury. Why should we care about group inequality? *Social philosophy and policy*, 5(1):249–271, 1987.
- [59] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.
- [60] P. R. Monge, P. S. Contractor, and N. S. Contractor. *Theories of communication networks*. Oxford University Press, USA, 2003.
- [61] Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6):066130, 2004.
- [62] L. Morrissey. Trolling is a art: Towards a schematic classification of intention in internet trolling. *Griffith Working Papers in Pragmatics and Intercultural Communications*, 3(2):75–82, 2010.

- [63] A. Mowshowitz. On the theory of virtual organization. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, 14(6):373–384, 1997.
- [64] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM, 2012.
- [65] 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.
- [66] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.
- [67] A. Parkhe, S. Wasserman, and D. A. Ralston. New frontiers in network theory development. *Academy of management Review*, 31(3):560–568, 2006.
- [68] W. G. Parrott. Emotions in social psychology: Key readings. 2000.
- [69] B. Pittel. On spreading a rumor. *SIAM Journal on Applied Mathematics*, 47(1):213–223, 1987.
- [70] P. Pomerantsev and M. Weiss. *The menace of unreality: How the Kremlin weaponizes information, culture and money*, volume 14.
- [71] A. Powell, G. Piccoli, and B. Ives. Virtual teams: a review of current literature and directions for future research. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 35(1):6–36, 2004.
- [72] S. Rayana and L. Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994. ACM, 2015.
- [73] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.
- [74] V. Sandulescu and M. Ester. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*, pages 971–976. ACM, 2015.

- [75] 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.
- [76] N. Shah, A. Beutel, B. Hooi, L. Akoglu, S. Gunnemann, D. Makhija, M. Kumar, and C. Faloutsos. Edgecentric: Anomaly detection in edge-attributed networks. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 327–334. IEEE, 2016.
- [77] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.
- [78] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
- [79] C. Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook. *BuzzFeed News*, 16, 2016.
- [80] E. A. Smith and R. Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14, 1967.
- [81] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.
- [82] J. E. Thomas. Statements of fact, statements of opinion, and the first amendment. *California Law Review*, 74(3):1001–1056, 1986.
- [83] M. Tomaiuolo. Trust management and delegation for the administration of web services. In *Organizational, legal, and technological dimensions of information system administration*, pages 18–37. IGI Global, 2014.
- [84] Q. Wang, W. Chen, and Y. Liang. The effects of social media on college students. 2011.
- [85] M. M. Wasko and S. Faraj. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pages 35–57, 2005.
- [86] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

- [87] A. Willmore. This analysis shows how viral fake election news stories outperformed real news on facebook.
- [88] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

Acknowledgments

My gratitude goes to everybody who has contributed to this work in different ways.

A special thanks goes to:

- My Ph.D. advisor Michele Tomaiuolo who has made many direct contributions to this work and for the interesting discussions that we have during this time.
- Agostino Poggi, Monica Mordonini, Stefano Cagnoni, Andrea Prati and Alberto Ferrari with whom I have worked in these years.
- Laura Sani, Paolo Fornacciari, and Gianfranco Lombardo who have shared their days and many brainstorming with me

First to my parents, which had to wait a long time for this achievement.

I also want to thank Susanna and Paolo, who, first of all, persuaded me to start publishing my works.

A special thanks goes to Mohamad and Kefah, wonderful couple of friends before fellow researchers.

To my school colleagues which worked strongly even in my place during these period.

To Matteo, Umberto e Francesco, great students whom I was lucky to work with in this three years.

Last, but not the least, to Marco, great friend and true leader, which always encouraged me to achieve my objectives.