POLITECNICO DI MILANO

FACOLTA' DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea Magistrale in Ingegneria Informatica

UNDERSTANDING THE DYNAMICS OF SOCIAL MEDIA INFLUENCE: EMPIRICAL ANALYSIS OF THE DETERMINANTS OF RETWEETING

Relatore: **Prof.ssa Chiara Francalanci**

Correlatore: **Ing. Leonardo Bruni**

Tesi di laurea di:

Matteo Freri - 784102

Anno accademico 2013 - 2014

A Valerio, Patrizia, Giuseppe

A Candida, Anna, Sara

Ringraziamenti

Un primo, sentito ringraziamento alla professoressa Chiara Francalanci, la quale mi ha consentito di partecipare a questo progetto di ricerca e mi ha supportato durante tutto lo sviluppo del lavoro con grande coinvolgimento.

Un grazie anche all'Ing. Leonardo Bruni, mio correlatore, per avermi seguito e indirizzato sin dall'inizio del percorso.

Grazie alla mia famiglia, che ha sempre assecondato le mie inclinazioni personali e mi è stata accanto anche nei momenti più difficili.

Grazie di cuore a Sara, che da anni mi supporta in ogni iniziativa ed è sempre in grado di regalarmi un sorriso.

Un grazie anche ai compagni di studi che hanno reso piacevole la mia carriera universitaria.

Grazie infine a tutti gli amici che mi hanno accompagnato o tuttora mi affiancano nel percorso della vita.

Matteo

Abstract

This graduate thesis is part of a wider research project whose aim is to analyze influence dynamics on social media. Literature focused its efforts into studying the role of influencers, i.e. users who have a broad audience (e.g. they possess many followers on Twitter). The term influence, instead, refers to the social impact of the content shared, regardless of who published it: the key point is the ability of the message subject to raise audience attention on its own. Though we are aware of the importance of being an influencer, our assertion is that message content possesses a decisive role in generating influence, irrespective of its author. Hypotheses testing is here performed with the aim of evaluating content significance in influence generation. An in-house software was built in order to support all the stages this thesis work consists of.

The first step assesses the weight of content specificity (i.e. level of detail) at user level, considering also tweeting volumes. Empirical results highlight a positive correlation between specificity and influence, while high volumes show a strongly negative connection with the possibility of being retweeted. Even when the popularity of the user is taken into account, specificity is shown to keep holding a positive effect over messages distribution.

The following section analyzes influence dynamics at single post level, without the bias of author variables: this is a crucial stage, where a clear distinction between influence and influencers is performed. Sentiment (i.e. feelings being conveyed) and specificity

are the employed variables. Data show a perfect fit to the model, validating the positive relationship between specificity and influence. As regards sentiment, the need of a few negative messages is displayed while seeking for a larger amount of retweets.

The final step of the work exploits data clustering, with the intention of verifying at which level specificity stops playing a significant role in influence generation.

Empirical findings are converted into guidelines, useful for both private users and corporations as a starting point for building a self-promoting strategy.

Riassunto

Questo lavoro di tesi si colloca in un più ampio progetto di ricerca, il cui scopo è analizzare le dinamiche della influence nei social media. La letteratura ha concentrato i propri sforzi sullo studio del ruolo degli influencer, ovvero degli utenti che possiedono un ampio pubblico (ad esempio, hanno molti follower su Twitter). Il termine influence, invece, si riferisce all'impatto sociale del contenuto condiviso, indipendentemente da chi l'abbia pubblicato: il punto cruciale è la capacità dell'argomento del messaggio di suscitare di per sé l'attenzione dei lettori. Anche se siamo consci dell'importanza di essere un influencer, la nostra convinzione è che il contenuto di un messaggio possegga un ruolo decisivo nella generazione di influence, a prescindere dal suo autore. In questo lavoro sono testate svariate ipotesi, allo scopo di stimare l'importanza del contenuto di un messaggio nello sviluppo di influence. Un software in-house è stato sviluppato per supportare tutte le fasi di cui è composta l'attività di tesi.

Il primo passo quantifica il peso della specificità del contenuto (cioè il livello di dettaglio) dalla prospettiva dell'utente, considerando anche i volumi di tweet. I risultati empirici evidenziano una correlazione positiva tra la specificità del contenuto e la influence, mentre volumi elevati ne mostrano una fortemente negativa con la possibilità di essere retwittati. Anche considerando la popolarità dell'utente, la specificità mantiene un effetto positivo sulla diffusione dei messaggi.

La sezione successiva del lavoro analizza le dinamiche della influence

a livello di singolo post, senza l'interferenza di variabili riferite all'autore: questo è un passo cruciale, dove viene operata una netta distinzione tra influence e influencer. Il "sentiment" (cioè le emozioni trasmesse) e la specificità sono le variabili considerate. I dati mostrano un perfetto adattamento al modello, convalidando la relazione positiva tra specificità e influence. Per quanto riguarda il sentiment, viene illustrata la necessità di pubblicare messaggi dal contenuto "negativo" per suscitare un maggior numero di retweet.

L'ultima parte del lavoro utilizza dei cluster per verificare a che livello la specificità del contenuto smetta di possedere un ruolo significativo nella generazione di influence.

Le conclusioni empiriche sono state convertite in linee guida, utili sia ai privati che alle aziende come punto di partenza per costruire una strategia promozionale.

Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Understanding – and consequently managing – the dynamics of information distribution has always been a tough task, especially on social media. People constantly share their opinions about everything they come in touch with: products, services, companies, places, and so on. This attitude implies a one-to-many relationship with other users, which possesses the ability to spread information in a very quick and broad way. Both companies and private users are interested in figuring out how these dynamics work, with the aim of increasing their influence (as an example, by deploying effective – though inexpensive – marketing campaigns).

This thesis work concentrates its analyses on the most known microblogging platform, called Twitter. This medium is ideal for studying both the information spreading dynamics and the importance of message content: everybody can take part into post sharing, even if he possesses no personal relationship with the author, and users can post short messages only (up to 140 characters), thus they need to concentrate on what they are saying way more than how they are conveying it [1]. Literature, up to now, has mostly concentrated its efforts in studying influencers, who are social media users with a broad audience (in this case, they possess a huge number of Twitter followers), while my work is mainly focused on analyzing influence, which is the impact that the message content has on the audience, regardless of the importance of the author. Even if the weight of being an influencer cannot be neglected, due to the broader reach that such a status grants, my

statement is that message content possesses a decisive role in generating influence, in spite of the centrality of the one who wrote the post in the first place. What is to be shown is the fact that content actually is a driver of social media influence and that variables such as posting volumes, specificity of content and message sentiment play a significant role in that.

Extensive hypotheses testing is performed both at author and single-post level, in order to assess the validity of the previous claims. The focus on the single post is critical, because it is the level where a clear distinction between influence and influencers is performed: while traditional media (such as radio or television) mainly rely on broadcasting communication, social media are populated with users who proactively share the content they are interested in, and it is not uncommon that influencers who post uninteresting messages are mostly ignored by their audience. Highlighting the importance of the content is the reason why it is so fundamental to analyze tweets at both levels, including user behavioral variables to the proposed models. The set of data employed in the analyses belongs to the tourism domain, which is one of the most active business fields on social media [2].

The following paragraphs describe the thesis general structure. At first, the state of the art is critically reviewed, then the formulation of research hypotheses at both levels follows. Afterward, the designed and implemented software tool is analyzed in detail. In the final section, empirical testing on the data is performed in order to assess what previously stated.

Chapter 2 illustrates the state of the art which refers to the core purposes of this research work. Both articles addressing the role of influencers and analyzing the relationship between influence and

influencers are discussed. According to the theory that content plays a significant role in influence generation, further related research works are explored. This thesis concentrates on the gap that exists on the impact of content over influence, both considering the user importance and independently from the centrality of the author.

Chapter 3 presents the research hypotheses, both at user and single-post level. As regards the user point-of-view, specificity of message content and tweeting volumes are taken into account as behavioral variables. For the first one, a positive correlation is supposed to hold towards influence dynamics, while volumes are believed to possess a negative relationship with the probability of being retweeted. Concerning the single-post approach, specificity and sentiment are considered. The former is still thought to have a positive effect over influence generation, whereas the latter ought to have a negative correlation. The cognitive process which led to the formulation of such hypotheses is here depicted: it took into account both the state of the art, my research team current work and some personal considerations.

Chapter 4 focuses on the description of the software tool which I developed from scratch with the purpose of supporting the statistical analyses. It is composed of four modules: the first one takes the dataset as input and consequently splits original posts from retweets; the following matches each original tweet with its respective retweets; the third module computes the numerical variables which describe a single tweet characteristics; the last one rolls up information on a user basis.

Chapter 5 is about hypotheses empirical testing. Different models are elaborated, both at user and single-post level, and the output computed by the software is given to them as input, in order to

either validate or confute the hypotheses. Control variables, useful for establishing the status of influencer of a user, are employed for assessing the value of content while the importance of a user is taken into account. The last section of the chapter makes use of data clustering at a single-post level for determining to which extent content keeps having a meaningful effect on influence. An effort is made to convert empirical findings into general guidelines, useful for a user or a company to promote themselves on the Internet.

Chapter 6, in conclusion, presents a summarizing discussion of the obtained results. In this same context, limitations of the performed analyses and desirable future additions to the current research work are illustrated.

Chapter 2 – State of the Art

This thesis work concentrates its analyses on the correlation between the content of a message and the social influence that it is able to generate. Such a relationship is investigated in depth, both at user level and single-post level (i.e., when the importance of the author is not taken into account).

This chapter introduces these topics and their context, while examining the related state of art. Section 2.1 frames the considered social media environment. Section 2.2 focuses on reviewing previous literature and stresses the difference between influence and influencers. Section 2.3 describes the recent change in perspective performed by research, which employs many different content-based variables in influence determination and has therefore been regarded as a basis for this work. Section 2.4 depicts the general theoretical model developed for understanding influence dynamics. Finally, section 2.5 recaps the gaps of current literature which are the focal point of this thesis.

2.1 The Context

The evolution of technology caused a worldwide spread of communication means that allow billions of people to be simultaneously connected even if they are miles away, creating an open virtual network. More and more individuals can afford either a laptop or a smartphone, which are low-cost devices useful for

connecting to the Internet, both through wi-fi nets and physical wiring. This has automatically brought to a truly social environment, where everybody can keep in touch with anyone in the community.

2.1.1 Web 2.0

Lately, these networks led to complex interaction systems that allow users to post their own content in great measure and to be proactive makers of what they are experiencing. These forms of communication are part of the technology widely known as Web 2.0 [94]: the overwhelming amount of information makes these websites – such as Facebook and Twitter – grow very fast and experience a huge traffic on a daily basis. The term "2.0" comes from the software implementation practice which uses such a notation to denote a newer version of a product: in this context, it underlines the evolution of the "old" Internet into a social environment.

One of the greatest changes implied into this transformation is the switch from static content pages to dynamic and interactive applications: exploiting new technologies, such as Ajax [95] and JavaScript [96], there is no need to refresh the page in order to view the updates. Older websites were filled with static pieces of information, written directly by the administrator of the website, and they had the main purpose of providing users with news about a certain topic. Though basing on the same network and protocols (like HTTP), this revolution allows the employment of databases, which can be used to store the data and can later be queried in order to gather dynamic information to fill the pages with.

In the Web 1.0 sites, users had a one-way read-only interaction with the content that the administrator published, while on Web 2.0 this

relationship is bidirectional and truly interactive: the growth of the website is mostly autonomous due to the great amount of data which users share; they are both listeners and speakers. This makes the quantity of information stored way greater for Web 2.0 websites with respect to static ones: if we suppose that every user is able to post the same amount of data that an administrator can, *n* different users would generate *n* times the data of an old website. The social interactions among them, in addition, would lead to a further increase of content. A negative aspect of this phenomenon is that the information overload that occurs does not allow the website administrator to check everything that is shared on his platform: it is mostly given to automatic tools and the community itself to spot and report inappropriate contents.

2.1.2 Social Media

The concept of social network roots back to more than a century ago [97], and refers to a group (of any dimension) of individuals interconnected by social bonds. These links can be of different kinds and with a different strength: we span from acquaintances to work colleagues, from friends to family. These relationships are moved to the virtual world and constitute the foundations of Web 2.0 communication means: the simplest way to depict them is to imagine a graph, in which people are the nodes and the edges represent the social interconnections. The more the relationships, the more the graph is dense (*refer to figure 2.1* [98]).

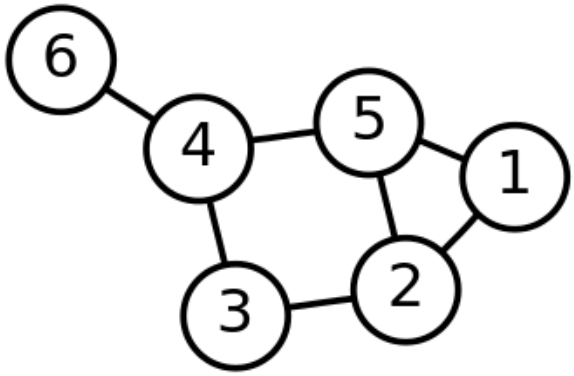

Figure 2.1: graph representing a social network composed of 6 people

Nowadays, social media [99] is a generic word referring to all the technologies that users can exploit for sharing their content online, mainly multimedial (photos, videos, …). They are a new way that people have for communicating and learning: everything they come in touch with is shared (think of Wikipedia [100], as an example). The variety of social relationships that people have is reflected in these new virtual means: Facebook [101], for instance, focuses on friendships, LinkedIn [102] on the job environment, Twitter [60] on the interaction between people with the same interests, and so on.

Traditional broadcasting media, like television or cinema, do not allow many people to publish their own content since they are very expensive: on the other hand, new social media let users share everything at low or none cost, and also corporations can perform effective – though inexpensive – marketing campaigns. A substantial feature that new media possess is the importance of the content: unlike traditional media, the subject of the communication is way more essential than the way it is conveyed (*refer to 2.1.3 and 2.3*).

Another important phenomenon which takes place on the social media environment is the one called "Wisdom of the Crowd" [103]: whenever a piece of information is shared, if nobody states that it is wrong, it is often assumed to be true, but one single denial can lead

to a mass rejection of the news. This attitude is in contrast with the typical "Suspension of consciousness" [73] which takes place in traditional media communication: people often believe whatever is told by television or newspapers, simply because they assume that there was a sources verification step prior to sharing anything.

2.1.3 Twitter

The Internet itself and the spread of Web 2.0 granted both people and corporations the possibility to share their own content, opinions and ideas in a way that makes them available to almost anyone [3]. Users actively exploit blogs, forums, social networks and other media with this purpose: sharing and giving feedback on what the others shared. This process is called e-WOM (electronic Word Of Mouth) and has been subject to growing attention throughout these years [4] [5].

The social media platform this thesis work concentrates on is called Twitter [59] [60]. Twitter is the most known microblogging [104] platform nowadays available on the Web: it allows users to post short messages (up to 140 characters) called "tweets", which can contain links to multimedia content and can be seen and republished – namely, "retwitted" – potentially by anyone in the community. This peculiarity allows an extremely quick diffusion of messages [6]. Unlike most social networks, Twitter relationships are based on interest, i.e. people "follow" the updates of a certain individual because they are interested in what he says. Peculiarly, connections are not bidirectional: someone can be followed by many other people but he may decide to follow a very small amount of other users. That situation is quite common with celebrities and companies, since they exploit this medium for marketing purposes and usually follow their

equals. Corporations take advantage of Twitter also for getting a lot of free – though not easy to collect – information on their users and products, with the aim of satisfying their customers wishes and increasing their profits. The overload of comments can be hard to manage and understanding which attitude companies should show on the Web in order to maximize their influence is a hard task, which is mainly addressed with this thesis work.

Other interesting features of Twitter are the one-to-many relationships which are established among users [7] and the need to summarize what has to be conveyed in a very short sentence: these characteristics make it one of the most suitable platforms for analyzing the importance of content in the online influence generation process. As already mentioned, another remarkable aspect is the concept of "follower". Everybody can see the tweets of a certain user without any sort of personal bond: celebrities, companies and associations have a public profile on Twitter and people can decide to follow their updates. The number of followers, which represents the width of a user's audience, is one of the most widely-recognized variables for identifying influencers (*see section 2.2 for further reference*).

Finally, Twitter offers the developers the possibility to exploit its public APIs. The dataset employed in this work is composed of tweets related to the tourism domain (which is one of the most complex and active in the social media field) which were collected through Twitter streaming APIs. These APIs came in use also later, when information about Twitter users (such as the number of followers) had to be collected.

2.1.4 Klout

Klout is another platform which was taken into account in this work [57] [58]. It is a website that tries to integrate all sort of social information about its users, linking their personal profiles from various networks (e.g. Facebook, LinkedIn, Wikipedia, Google+, Blogger, …). Its motto is "Klout: the standard for influence" and it summarizes a user online reputation through a numerical value called "Klout score". Such a score is continuously updated and its value keeps changing day by day, accordingly to the user activity on the social media.

Actually, section 4.2.7 of this thesis empirically hints how the Klout score is not really representative of the real user influence: it is, instead, a proper index for measuring the status of influencer (i.e. user with a broad audience; *see 2.2*), since its computation is heavily biased by the user number of Twitter followers. That is also because Twitter is one of the most open social networks, and allows to retrieve all sort of information on their users even without their permission, while others do not. Other works have addressed this issue and supported this hypothesis, showing how the Klout Score is not really a measure of influence as we are defining it [52] [105].

Exploiting its APIs, I was able to collect the Klout score value for the Twitter users involved in the considered dataset. Klout score was therefore used in this work along with the number of followers as a control variable representing the importance of a user.

2.2 Influence Versus Influencers

Literature on social media makes a difference between the concepts

of influencers and influence. Previous research work have mainly focused on studying the role of influencers, i.e. users who have a broad audience. They either possess many followers on Twitter, or have a lot of friends on Facebook, or are included in many Google+ circles, and so on. The term influence, on the other hand, refers to the social impact of the content shared, regardless of the user who published it in the first place.

Traditional media, such as television and press, have identified the breadth of audience as the primary index of their influence (just think of Auditel, as an example), and following this idea many research works have recognized social media prominent users as the most influential online sources. As already said, the centrality of a user is one of the main variables considered while evaluating the importance of an individual in a network, as that is an inborn feature of any network-like system. Linton Freeman, professor at University of California, introduced the first centrality metric, namely degree centrality, in 1979 [8]. That is an easy-to-compute number, which represents the amount of links that a node possesses (or, if normalized, the percentage of links over the whole network). Its value is straightforwardly understandable: if a node is connected to many others, it is expected to play an important role [9] [10]; it was also shown how a high value of degree centrality is usually connected to more active participation into network activities [11]. The relationships among network nodes may, in addition, be asymmetric [12]: in the case of Twitter, people can follow another user, but they may not be followed in return (as mostly happens e.g. between fans and celebrities). This difference between followers (i.e. inbound links or "indegree centrality") and followees (i.e. outbound links or "outdegree centrality") allows users to receive updates without the need of knowing each other in advance. The number of followers, i.e. the indegree centrality, is the most typical index used

for identifying social influencers on Twitter.

Other researches have compared a user leadership with the authority of a website, analyzing it with the same metrics. The importance of a node is consequently related to the importance of those who are linked with it: the more the followers have weight, the more the user has. A very well-known index, which reflects such a measurement procedure, is the PageRank. It is a score invented by Google creators Page and Brin, in 1998, frequently adopted to evaluate influencers because it is theoretically similar to degree centrality [13]. Researches revealed that an author with a higher PageRank is likely to have a larger tweet propagation than a user who possesses a lower value, and that he usually has a higher number of followers (i.e., both are metrics suitable for determining whether a user is an influencer or not) [14] [15]. Nevertheless, in other studies the PageRank was shown to be significantly uncorrelated to the number of retweets that the user received [15] [16].

These recent works (2010) have given birth to a stream of literature which does not consider degree centrality metrics only while attributing to a user the status of influencer, but also other behavioral variables. In 2009, they were preceded by a small study which analyzed twelve Twitter users in-depth, highlighting how the number of their retweets (i.e. a simplified metric of influence) was not only related to their number followers but also to their own tweeting volumes [17]. Even tough the dimension of this study is too reduced to allow generalization, it was one of the first to promote a shift of the attention of researchers to different metrics. Another research showed how, tough the number of followers can be significant in generating users feedback for some categories of accounts (such as news providers), the importance of other metrics (like the number of mentions) cannot be neglected [18]. Finally, in

2010, a large-scale empirical study was performed: on a dataset composed of six million Twitter users, no significant statistical correlation between the number of retweets and mentions and the number followers was found [19]. The inbuilt complexity of the influence phenomenon led the authors to conclude that the identification of true influencers should not be based on a single metric only.

2.3 A Change in Perspective

The weakness of the aforementioned works is that they do not take into account the fact that traditional media are focused on broadcasting and do not allow interactive communication with their consumers. In this different social context, being well-known is not enough: everyday evidence shows how influencers who publish uninteresting content are mostly ignored and how, in contrast, ordinary users manage to achieve a good feedback if they post appealing material [19] [20].

Recent works on social media, aware of their unique features, have linked the notion of influence with the content of the message being shared [4] [7] [21] [22] [23] [24]. Some other research projects have also highlighted the importance of "ordinary influencers", i.e. everyday users who have a strong impact on a small fraction of followers (like one or two) [4] [25]: they are more cost-effective than well-known influencers and maximize marketing efficiency [7].

It is quite complex to make everybody agree on what influence actually is, therefore quantifying it can be a tough task. Obviously, this requires a major change in perspective: evaluating influence does not provide a static ranking of influencers anymore, since it is

based on the many different aspects that can be considered while outlining the content of a message.

2.3.1 Syntactic Variables

A great variety of content features has been taken into account in latest literature, with the purpose of delineating further metrics useful for spotting real influencers. The presence of hypertext links inside a tweet has been regarded from the start as primary, and many different studies assessed its value in increasing the retweeting probability and in helping researchers performing topic adoption analyses [4] [22] [26] [27] [28] [29] [30]. The presence of trending words, prefixed with the "#" symbol (namely "hashtags", in Twitter context), was shown to be useful as well in rising audience attention [22] [27] [31] and in constructing models for identifying the topics of the shared content [28] [32] [33]. Hashtags came also in hand in content-based filtration [34] and in users portraying [35]. Another content-related variable which was analyzed and confirmed to possess a positive effect over influence generation is the inclusion of mentions (i.e. references to other users) [27]. Models predicting the presence of mentions were made available through the employment of hashtags and links themselves [26]. An additional research work (which considered both Twitter and Facebook users) highlighted the importance, for the so-called influencers, of interactively engaging their audience with original content in order to have an actual impact [36]. Finally, another study has concentrated on distinguishing between through self-promoting messages and pure information sharing tweets using the categorization of content (namely, the topics) with the aim of outlining in a more accurate way the dynamics of the retweeting process [37].

In this thesis work, when I talk about content specificity I refer to a variable which takes into account both hashtags, words and links for classifying a message into a set of predefined subject categories: that will be one of the variables essential in determining content effect over influence generation.

2.3.2 Semantic Variables

Sentiment analysis, which is the use of natural language processing for extracting subjective information (i.e. opinions and feelings) from a message [56], can be exploited for improving the accuracy of the models that either employ degree centrality metrics or make also use of content-derived syntactic variables. Plenty of works were written of this topic, because opinion classification (OC) has always been of great interest for the scientific community [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48]. These researches analyze the text both at document level and at single-sentence level (in the case of tweets, these two aspects practically coincide). The former process produces an overall document evaluation, consequently classifying its general feeling as either positive, neutral or negative. The latter is usually considered a harder problem and aims at classifying sentences as either subjective or objective (i.e. carrying sentiment or not). As shown both in [49] and [50], sentiment evaluation is not an easy task, and its results are highly domain-dependant.

The main purpose of the aforementioned works is an aggregate estimation of sentiment at different levels. Since users usually have various opinions about different aspects of the same fact, such an approach might not be detailed enough in some analyses. As an example, distinct features of a product may imply dissimilar reactions [51]: a consumer could appreciate the interactivity of a smartphone

very much, but he may not be glad about its low battery duration. The result would be a positive overall polarity of the message, hiding the negative opinion on that specific characteristic. Feature based sentiment analysis has the purpose of identifying the traits of the subject the user is talking about (i.e. feature extraction, FE) and assigning a different polarity value to each of them.

This thesis work combines both the described approaches, namely opinion classification and feature based sentiment analysis. The dataset in use, collected by my same research team, contains many rows for every tweet, each of them expressing a distinct sentiment value – positive, negative, neutral – for one of the characteristics of the message which carries a feeling. In order to have a single variable able to express the sentiment without losing all the information carried by feature based analysis, I assigned a numerical value to each sentiment symbol (-1 to the negative, $+1$ to the positive and 0 to the neutral) and made an average of these values for the single tweet (*refer to section 4.2.5 for further details on these computations*). The result is a real number, spanning from -1 to $+1$, whose plus (minus) sign indicates the prevalence of positive (negative) feeling in the post – with 0 meaning overall neutrality – and whose value represents how either positively or negatively biased is the opinion of the writer (e.g. -0.8 means mostly negative, $+0.1$ slightly positive). This allowed the data to perfectly fit the single-post level model, which makes use of the sentiment variable.

2.4 A Wide-Ranging Model

Up to know, I have discussed which behavioral variables can be employed for determining the importance of a message and how

they impact over influence generation. What is left to do is deciding how to concretely quantify influence (i.e. which variables to use).

As regards Twitter, previous research works pointed out how the influence of a person is correlated to the width of the feedback received, that is the number of retweets [19]. Literature has also acknowledged the importance for a user to see his content shared frequently, awarding time dynamics a valuable weight [52] [53]. Consequently, three variables have been employed in this work for measuring influence, namely the retweeting frequency (which is the percentage of the tweets of a user which have been retweeted at least once), the amplitude (the average number of retweets received) and the persistence (a more complex variable, taking into account different aspects of time dynamics) (*see 4.2.4 and 5.1.1 for further details on the computation of these variables*). Figure 2.2 represents the conceptual framework this work is based on: it covers all the previously-described aspects of the phenomenon, from the causes to the effects. The different levels of aggregations are highlighted and the practical employed variables are shown in the lower boxes.

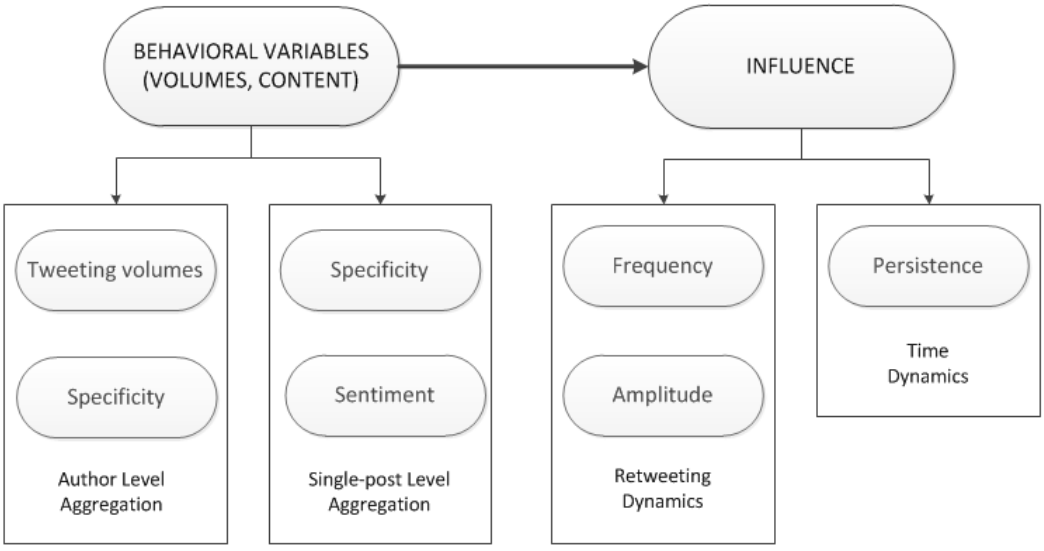

Figure 2.2: overall conceptual and practical framework

2.5 Literature Gap

This State of the Art chapter was meant to review literature on social media, especially the articles regarding the difference between influence and influencers. As already discussed, researches up to now mostly concentrated their efforts in analyzing the role of influencers as users with a broad audience, while the concept of influence is somewhat unexplored.

My thesis work focuses on using content-based variables derived from the behavioral decisions made by the user (such as the tweet sentiment, the posting volumes and the specificity of the content) for studying influence dynamics. Both sentiment and specificity are computed in an innovative way, which makes them variables spanning on real values and allows their significant contribution to statistical models (*refer to 4.2.5, 5.1.2 and 5.2.2 for further information*). The evaluation of influence dynamics is performed both on a user and single-post level, as shown in the conceptual model presented in the previous section. These analyses allow the formulation of general behavioral guidelines, which can be used both by companies and private users with the aim of increasing their online influence and creating concrete business value. Such principles are awarded great importance among corporations, which are very interested in deploying cost-effective marketing campaigns [54] and need to count on reliable guidelines [55].

Chapter 3 – Research Hypotheses

The hypotheses verification work described in this thesis is divided in two parts: the former analyzes the retweeting dynamics at user level (retweeting process model "in-the-large"), while the latter considers the same dynamics but on a single-post perspective (retweeting process model "in-the-small") (*refer to paragraph 4.2.4 for further details on these two modeling perspectives*). This chapter is devoted to describing the theoretical process which led to the formulation of the research hypotheses on both perspectives. The empirical testing, instead, is described in Chapter 5, after the portrayal of the implemented software tool: it makes use of a dataset consisting of slightly more than a million and a half tweets, taken from the tourism domain, spanning over a three months period.

As regards user level dynamics, some preliminary analyses were shown in [61] [62] by my same research team: considering that as a starting point, this thesis work formulates new hypotheses about importance of content over influence, tests them on a model at the user level viewpoint and makes numerical comparisons and critiques. Additional effort was put into showing how such theories hold, though – of course – on a smaller extent, while the popularity of the user is taken into account (i.e. considering metrics indicating that a user is an influencer, such as the number of Twitter followers or the user Klout score).

Regarding single-post level dynamics, I formulated three original different hypotheses (binding behavioral variables and influence),

showed the realization process of a new model which allowed their testing and consequently checked their validity. In the empirical testing section (*see 5.2.3*), I also make use of some clusters derived from the dataset, in order to highlight some characteristics of the tourism domain being considered.

The final goal of the entire analysis, at both levels, is to translate these findings into guidelines, such as which frequency of posting should be adopted or what level of specificity of content should be reached, in order for authors to maximize their influence. On a theoretical point of view, these empirical evidences raise additional challenges and encourage extra research to better understand the relationship between content and influence on social media.

3.1 User Perspective

As regards the retweeting dynamics which take place at user level, two author behavioral variables were taken into account, namely volumes and specificity (*refer to the State of the Art chapter for this choice, 2.2*), whose correlation with influence was then investigated. The first of the pair, volumes, is a quantitative measure which corresponds to the amount of messages shared by the user (in Twitter context, tweets). The second variable, instead, symbolizes the level of detail that an author reaches while writing on a specific subject. Making a comparison with the context of public speaking, literature has highlighted many times how such variables are one of the main concerns for lecturers, either they are general purpose orators [63] [64] or teachers [65]. The situation in the social media environment is alike, as users have to correctly choose how much to post and how specific to be in their messages, in order to achieve the best possible outcome.

Principles leading public speech dynamics were accepted here as a general guideline: these principles, widely recognized as effective in impressing the audience, consist in making short speeches, if possible along with visual content and immediate catchphrases [66], and engaging the listeners in a conversation, in order to both verify their understanding and repeat, reinforce and support the point being made [63]. In such a context the audience is, most of the times, not related to the speaker by any sort of bond and may not have information about him: that disregards one of the distinctive features of social media sites, namely that two users are usually tied with some sort of social relationship (at very different levels of depth and strength, depending on the platform itself [67]), but that is acceptable since the social connection between author and retweeter is not subject to investigation in this work and is often not present on this particular social network.

As a matter of fact, the main goal of a social media user who wants to be active and influential is to post something that is shared frequently, by many other users and over extended periods of time before fading [52] [53]. The three variables which represent these peculiar features are namely retweeting frequency, amplitude and persistence (*refer to 4.2.4 and 5.1.1 for their computation details*): these are employed in my framework model as main drivers of influence (*look back at section 2.4*). The importance of the correct control over these variables is easily recognizable: everybody has experienced, at least once, the uneasiness of reading the posts of someone who either keeps updating his profile too often or publishes content which is of no interest at all – even annoying (as an example, think of a person on Facebook who keeps filling your home or wall with political messages which you do not agree with; he is likely going to be unfriend sooner or later) [14] [23] [68]. On the

other hand, it has been shown how users not contributing to social conversations are almost irrelevant to determine the dynamics of influence [19]: though there is no recognized best practice, since the posting volumes of authors heavily vary from one to another, a balance of the two is supposed to be the best option. Literature has not provided a methodical evidence on how the behavioral decisions illustrated above – about volumes and specificity – impact influence: the model taken into account and the hypotheses verification correlated to it (*see 5.1.1 and 5.1.2*) are a step forward this direction, and have the goal of concretizing the wide-ranging model (*back in section 2.4*) describing the relationship between content and influence.

As regards the hypotheses, a first proposition was developed considering two different needs, one inborn into public speaking and the other related to social media distinctive dynamics. The former consists in conveying general messages, in order to achieve a greater impact on the listeners: as widely proved in literature, giving a summary of what is being presented, reiterating the point and avoiding details (though often they are required) helps the audience to recall the speech more easily [63] [66] [69]. The latter is associated to social media "long-tail effect": a great variety of small communities, very interested in specific content, populates social media websites, and the peculiar organization of each of these groups characterizes a different type of media [70]. Due to fact that, in these communities (especially on Twitter), the relationship between orator and audience is built around the interest into a specific subject, we can suppose that specificity no longer has a negative contribution to the addressees' attention and response, differently from what literature has said about public speaking. This is the way in which the distinctive features of social media are considered predominant over the general principles enounced before.

Proposition 1: Specificity of content is positively associated with influence.

This same statement was made concrete through the employment of the three abovementioned influence metrics, namely frequency, amplitude and persistence of the retweeting process (*refer to 4.2.4 and 5.1.1 for their computation details*), which allowed the formulation of three corresponding hypotheses, which tie specificity and influence.

Hypothesis A: Specificity of content is positively associated with frequency of sharing.

Hypothesis B: Specificity of content is positively associated with amplitude of sharing.

Hypothesis C: Specificity of content is positively associated with persistence of sharing.

While the effect that specificity is supposed to possess over influence generation is uniform (namely, positive), volumes are thought to bare contrasting effects.

According to the reasoning that a user that posts too much can become either annoying or boring to his audience [14] [23] [68] [71], high amount of tweeting should correspond to overall lower retweeting frequency (which is the percentage of retweeted tweets over the total posted ones), since it measures the level of engagement of the audience that the author was able to reach. Marketing research, in contrast, has always suggested that repeating an advertising message multiple times (i.e. increasing its volumes) is

beneficial to the effectiveness of a campaign [72]. Though that mainly applies to traditional broadcasting media – where the phenomenon of "suspension of consciousness" takes place and reduces the importance of the content itself [73] –, such an effect cannot be neglected completely: once the attention of some readers has been grasped and consolidated, a greater quantity of tweets can help maintaining audience consideration through time and expanding market share. These considerations let the following hypotheses be formulated:

Hypothesis D: Volumes of content are negatively associated with frequency of sharing.

Hypothesis E: Volumes of content are positively associated with amplitude of sharing.

Hypothesis F: Volumes of content are positively associated with persistence of sharing.

3.2 Single-Post Perspective

Though I am aware that being a well-known individual or corporation, with established reputation, is a factor that has major impact over retweeting dynamics (*as shown both in literature, section 2.2, and in the empirical testing section, 5.2.2*), I wanted to illustrate how properly selecting tweeting variables, such as the aforementioned specificity of content or message sentiment, can improve a post visibility and influence, regardless of the one who posted it in the first place. As an example, we can consider the "network phenomena", people who are recognized influencers, but used to be almost-unknown youngsters that had something very

meaningful to say: what had impact on their audience – and allowed them to increase it to the extent it has now – was the essence of their messages, and not their popularity [52].

In this different context, the behavioral variables which were taken into account are specificity and sentiment (*refer to the State of the Art chapter for this choice, 2.2 and 2.3*). At single post level, "volumes" variable loses any meaning, since it assumes value 1 for each considered tweet, so it is discarded. In order for the considered data to have a chance to fit the model, only original tweets with at least one retweet were taken into account: if also tweets without any retweet were considered, they would possess no metrics for determining the generated influence, because we are not at author level any more and only one tweet is considered at a time. Frequency, in this new context, becomes meaningless too, since it represents the number of tweets which have been replied for each tweet considered, i.e. it is 100% for each single post. Amplitude can still be measured, but it is basically equal to the number of retweets the post received. Sentiment variable – namely positive, neutral or negative, depending on the feeling the author of the message wanted to convey – was added to allow a deeper understanding of the retweeting dynamics, in spite of the importance of the user, because it has always been regarded in literature as a major element in determining the effectiveness of communication (*see Chapter 2.2.3 for the references to works related to sentiment analysis*).

Two propositions, named 2 and 3, (*to distinguish them from the previous one – see paragraph 3.1*) were developed and consequently split into four different hypotheses, which concretized them into specific variables relationships. Proposition 2, based both on literature ([63] [64] [65] [66] [69] [70]) and, in retrospect, on what was empirically proven for the previous model (*refer to 5.1.2 for the*

empirical assessment), states the same thing proposition 1 suggested, which is

Proposition 2: Specificity of content is positively associated with influence.

Proposition 3, instead, is innovative, since it is about the sentiment of the content being posted. Literature on social media provides evidence that users have a general tendency of self-promotion and generate most messages with positive sentiment [74] [75], while, on the other hand, traditional media usually highlight negative news [76]. As proven into [61], [77] and [78], though the majority of tweets carries a positive sentiment, those which trigger the biggest retweeting phenomenon are negative ones. Regarding persistence, even if data distribution seemed to indicate a tendency towards negative tweets being retweeted more quickly, there was yet no evidence for such a relationship. Consequently, the proposition can be formulated as follows:

Proposition 3: Sentiment of content is negatively associated with influence.

That sentence means that posts carrying a negative feeling have a greater chance of being retweeted and a general tendency to persist more (i.e. being retweeted more quickly but lasting longer in time): while the first idea has already found an empirical validation, the second one is still to be tested.

Through the concretization made possible with the employment of the influence variables, these were the four generated hypotheses (enumerating letters follow the previously used ones):

Hypothesis G: Specificity of content is positively associated with the number of retweets.

Hypothesis H: Specificity of content is positively associated with persistence of sharing.

Hypothesis I: Sentiment of content is negatively associated with the number of retweets.

Hypothesis J: Sentiment of content is negatively associated with persistence of sharing.

Chapter 4 – Analysis Tool

As already mentioned, this thesis work is part of a wider research project whose main aim is studying influence on social media. The term influence is here used meaning the social impact that the content of a user's message is able to carry, regardless of the importance of the author. Understanding these dynamics requires an automated software that can manage great amounts of data and allow the requested analyses.

4.1 Overview

This chapter focuses on describing the architecture of the in-house software tool and the conceptual steps behind the design choices, justifying the implementation decisions taken. Initially, the provided dataset is described in detail, framing the analysis domain. The first software module, namely "Originals and Retweets Identifier", is depicted in the following section and aims at splitting the source dataset into two distinct sets, one composed of original posts and the other of retweets. The second module, "Tweets and retweets matcher", matches each original tweet with its respective retweets. Before proceeding to the third module ("Data summarizer"), which computes the numerical variables describing the characteristics of a single tweet, an overview on the concepts behind the retweeting process modeling employed is presented. Other two steps are required before the execution of the last system module can take place: an evaluation of the retweeting queue dynamics (in order to

increase the statistical significance of the obtained results) and an empirical assessment of Klout score as an index of the influencer status. The last tool module rolls up the information gathered on a user basis and gets new data about each user. Picture 4.1 follows this stream of analysis and shows the entire software architecture, with the most meaningful steps highlighted.

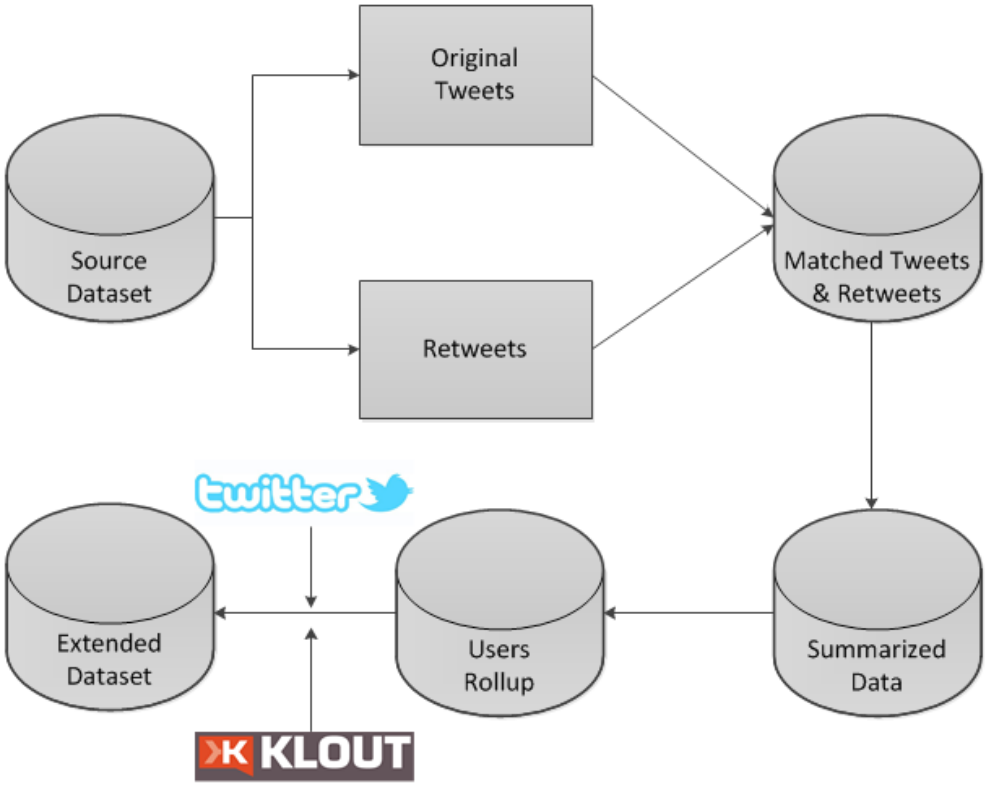

Figure 4.1: high-level software architecture

4.2 Design and Implementation Steps

The proposed architecture was implemented in Java (version 7 SDK), using Eclipse Indigo for Microsoft Windows as IDE, and exploits Twitter4j [79] and Klout4j [80], which are unofficial Java libraries useful to interact with Twitter and Klout APIs. These two libraries

show a few limitations (such as little error codes and messages displaying or missing support for some functionalities), but are generally practical and performing and excellently fulfilled their tasks in this thesis work. MySQL 5.5 is the RDBMS used to manage the proposed dataset and to both save and query all the new data generated by the software tool itself. Whenever possible, all the queries performed on database tables were optimized through the use of indexes (B-Trees or Hashes). As a side note, tweets and retweets may be generally referred to as "posts" in the following sections.

$4.2.1$ Provided Dataset

The dataset I have been working on was provided to me by the information systems research team of Politecnico di Milano, which has been developing studies about Twitter influence for a long time. It is composed of one single table, filled with raw tweets collected through Twitter streaming APIs, by means of an automated ad-hoc tool. Tweets were collected from the 1^{st} of December 2012 to the 28^{th} of February 2013, for a total amount of $1'511'497$ different posts. The analysis domain is tourism, thus all the queries performed to Twitter contained one of the following crawling keywords, representing some of the most famous Italian tourist destinations: Amalfi (and Amalfi Coast), Lecce, Lucca, Naples, Palermo and Rome. Every stored post has an attribute "brand" in the database that identifies which city is the tweet about (*refer to the "Anholt Nation Brand index model" classification just below for further details*). English and Italian are the two languages that were considered in the collecting process.

Every tweet is enriched with some piece of information, such as id of

the post, language, author name and id, publication date, sentiment of the content and subject categories. Each post in the dataset is uniquely identified by the pair id-language, namely "post_id" and "cdb_language_id" in database attribute names, since a single "post_id" may be repeated for different language ids (e.g. pairs of "post_id" $= 151'674$ / "language_id" $= 15$ – English – and "post_id" $= 151'674$ / "language_id" $= 52$ – Italian – can coexist, indicating two completely different tweets). Each post was assigned to a "brand" (i.e. city) and to zero or more categories by a semantic text processing engine [81], which has been instructed to classify through a different network of keywords for each considered category. The set of categories was derived from a modified version of the Anholt Nation Brand index model [82]; examples of such categories are "Arts and culture", "Events and sport", "Food and drink" and "Fashion and shopping", for a total of 44 different kinds [62].

4.2.2 Original Tweets and Retweets Identifier

The first section of the developed software is meant to split the provided source dataset into two disjoint sets, one containing all the original tweets written by users and the other one with all sorts of retweets.

According to [83], retweets are text messages identified by one of these expressions: 'RT @', 'via @', 'MT @' or '"@' (followed by the username of the person whose tweet is being quoted). The software identifies if a tweet contains one of those terms and, in case it does, consequently classifies it as a retweet. Though Twitter has a "Retweet" button, that by design adds the exact expression 'RT @' (space included) to the new user message while performing a retweeting action, I took into account that many people write their

replies by hand and consequently may have omitted the space or composed their messages lowercase: these checks increase the overall accuracy of the classification.

The identifier does not distinguish between first time retweets or posts that have been retweeted many times in a row: as an example, "RT @Marcus RT @John Hello" means that a third user retweeted Marcus, who had retweeted John in the first place. These are altogether considered retweets, no matter how long the retweet chain is, otherwise influence would limit itself to direct retweeters, losing its authentic meaning. The issue of "cut&paste retweets", namely those retweets which are an exact copy of another tweet – containing none of the typical retweeting expressions –, was not considered, since there is no actual way to distinguish whether one of those is a copy or an original post. Original tweets are text messages without any sign of retweeting activity.

When the program is executed, it requires the user to insert the number of database tables the dataset consists of. Though it is possible to insert any positive integer number, the tool is optimized to work on one table at a time, so it would be better to condensate all the available data in a single one, in order to obtain a much faster execution. Following the given dataset conventions, different tables must be marked with a unique identifier (attribute "job_id" in the database schema): the software asks the user what "job_id"(s) it should work on, thus making the condensation activity not an issue for the analysis itself, because everything can be merged in a single table but distinguished via "job_id". The tables the tool works on are called "snippet_"+"job_id" (e.g. "snippet_279").

Since the original dataset can be composed of an arbitrarily large number of tweets, the tool does not load all of them at once into the

program heap memory: such an action could fill up the available space very quickly, causing the Java Virtual Machine to crash, especially with huge datasets [84]. Consequently, posts are queried by groups of 30'000 (this number could be somehow different, but it was chosen empirically, in order to maintain a balance between memory occupation and efficiency), ordered by "post_id" and "cdb_language_id", then they are classified and saved into the two respective database tables (straightforwardly called "originals" and "retweets"), allowing heap memory to be freed before the following fetch. Whenever it is time to gather the following group of tweets, the tool recovers the "post_id" of the last tweet fetched in the previous round and starts querying the next group setting the "post_id" attribute equal or greater to that one. The "cdb_language_id"s of such "post_id", which were already taken in the previous round, are also memorized: this way, in case only part of the "post_id" – "cdb_language_id" pairs were fetched and classified, it is possible to know which ones are new and which ones are to be discarded, because already taken into account. As a clarifying example, imagine that the $30'000^{th}$ tweet fetched in a round has "post_id" $= 123'456$ and "language_id" $= 15$. In the following round, the query gathers posts starting from "post_id" $= 123'456$, but is able to discard the one with "language_id" $= 15$ because already fetched, and can consider the one with "language_id" $= 52$ only, because it has not been classified yet.

This whole process is necessary due to the SQL "limit" keyword, which I used in my queries for fetching a limited amount of rows (in this case, $30'000$ per time): its syntax is "limit X,Y", where X and Y are lower and upper row numbers (e.g. "limit $0, 30'000$" gathers rows from 0 to $29'999$; "limit $100, 700$" from 100 to 699) [85]. The issue is that it needs to go through the whole table every time in order to find the lower row limit [86]: due to potential different size and

deletions of rows, it cannot just jump to the right row directly. This makes an execution slower and slower as the block analysis proceeds, because the lower limit keeps increasing (e.g. it would take a long time to reach the lower row on the 101^{st} group, with "limit $3'000'000, 3'030'000$"). The easiest way I found was, therefore, to use "limit $0, 30'000$" in each query (making it instant to start from the first row available, which is always 0), but attaching the clause "$WHERE\ post\_id\ >= X$", with X equal to the last "post_id" fetched, which filters out posts already fetched before the limit is applied. This allows the query to consider the remaining posts only, selecting the 30'000 needed way quicker.

The program automatically determines the temporal extension of the dataset provided (from/to which date its tweets go) and shows it to the user, who can decide whether to perform the analysis on the whole period or select a shorter interval of dates, thus ignoring the tweets which do not fall into that time span.

The system can also filter, through keywords, the tweets to perform the analysis on: as an example, if a user wants to work on tweets containing "Rome" or "Naples" only, he can type those two words when prompted, and the software will discard the tweets which do not contain either "Rome" or "Naples" while fetching from the database. A filtering function matching all the keywords inserted was not implemented, since it is pretty rare to find single tweets containing more than one exact word: looking for posts containing both "Rome" and "Naples" (if any) would end into a too small set of results, thus making it useless for the analysis itself.

The way the system is designed allows the user to stop the execution of the software anytime during the fetch of the posts, resuming it later from the last "post_id" it was interrupted at: this is

useful in case the analysis takes too long or a break is needed, for any reason. When the application is restarted, it is possible to erase all the data from the tables being filled ("originals" and "retweets"), in case any error might have occurred (e.g. power supply off, hardware or software failures, …) or if it is necessary to perform another analysis on a different or wider dataset.

Overall, the originals-retweets split took about three hours on the current dataset (approximately one minute every $10'000$ posts), which resided – as already mentioned – in a single table.

4.2.3 Tweets and Retweets Matcher

Whenever the previous phase terminates, the software waits for user input in order to proceed to the next part (that is the default behavior for each step).

This second phase works with the assumption that originals and retweets tables are both generated by the tool itself, thus the identifiers of their tuples are contiguous and self-incrementing, because that is how the software generates them. Queries performed will consequently contain an order-by statement, having the row id field as an argument. This allows the analysis to be stopped and resumed later, starting from the first post whose id has not been checked yet.

At first, the tool fetches a group of original tweets from their table (specifically $100'000$ per time, balancing memory occupation and program efficiency, as usual), then starts getting one retweet per time from the other table and tries to match it with any of the original tweets currently fetched. For every group of original tweets

fetched, the whole retweet table is scanned, but any time a retweet corresponds to the original post it comes from it is marked as "matched" in the database. This way, while the software proceeds in the analysis, the number of retweets to check will be smaller and smaller (already assigned ones will simply be ignored by the query) and each round will be quicker than the previous one. On the current dataset, the matching time lowered from forty-five minutes of the first round to fifteen minutes of the last one, for a total amount of about four hours. Overall, this method is way faster than getting each original post and then look for its retweets: that would mean scanning the whole retweet table for every single original (since a tweet can have multiple retweets), causing the matching time to be way longer than an acceptable amount. The current method, instead, is very quick: the scanning stops whenever a retweet finds its own original post, since it can be one only. All the matches between a retweet and its original were saved in a new database table called "ajob", which contains one row for each pair found.

The matching between a retweet and its original post undergoes the following rules: first of all, the retweet must contain the exact text the original message had, and their languages must be the same. It is also important to take into account the date in which the post was published: an original tweet must have been written before any retweet action was performed. Last, but not least, the retweet must contain the name of the author of the original post, preceded by one of the typical retweeting expressions mentioned before [87]. In case a post overcomes all these checks, the assignment takes place and the tool can switch to the following retweet.

On the current dataset, about 80% of the retweets matched with their respective original post: due to Twitter APIs, that filter which tweets to return to the querying user [88], the full set of posts might

not be available, and we must also take into account that some originals may belong to a time period preceding the one considered by the dataset (and obviously the tool is not able to relate retweets to originals which were not retrieved). In spite of all this, the matching percentage is quite high and allowed the requested analyses.

At the end of this phase, after all the retweets have been assigned to their original tweets, the software checks if their text contains a link and consequently marks them in case it does. The search is performed looking for either "http://" or "https://" keywords in the text, since most of the tweets do not contain normal links but short-links only, which do not have any "www" or typical domain extensions (an example of short-link is "http://t.co/12345", and nowadays - most of the times - Twitter automatically reduces the posted links to such compressed standard format [89]). I personally checked that, on the current dataset, all the Twitter t.co short-links start with either "http://" or "https://" (I looked for posts containing ".co" only, excluding those with ".com", "http://" and "https://", and none were found), thus I can state that the search keys can be considered pretty accurate in link finding. Such knowledge about links contained in tweets was not specifically brought into play in the current model identification, but this tool is going to be used by the research group to perform further analyses, so I tried to make it as flexible and useful as I could, in order to allow a wider spectrum of investigation possibilities. This is the reason why I synthesized the most information about tweets that I was able to, such as the number of hashtags contained in each post or the ratio between a single tweet length and the overall average one (*refer to section 4.2.5 for further details concerning these additional computations*).

4.2.4 Retweeting Process Modeling

Here follows an explanation of the modeling performed in order to overcome the inherent complexity of the retweeting process. Such a procedure was mostly developed by the same research team that provided me the dataset [62] [81]. This is a compulsory step for understanding some of the database attributes synthesized in the following phase.

The process can be split into a "retweeting process in-the-large", that focuses on the retweeting dynamics at user level, and into a "retweeting process in-the-small", which instead concentrates on the individual tweet level. We denote with T the total observation interval in which all users' actions take place.

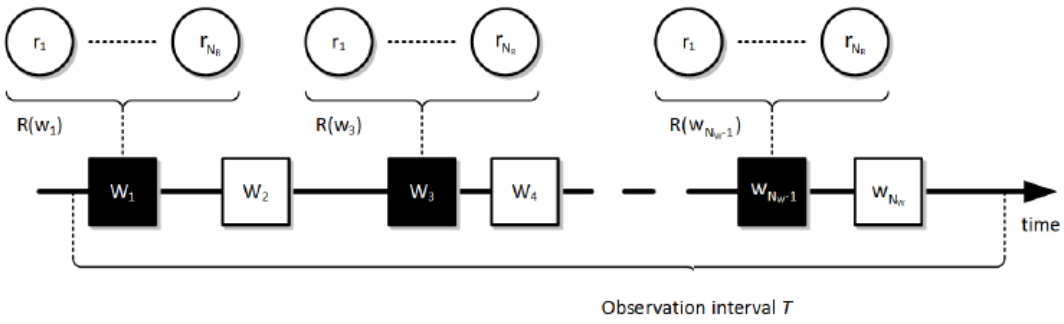

Figure 4.2: "in-the-large" retweeting process model dynamics

As regards the retweeting process in-the-large, we consider a generic Twitter user (referred to as X) and his posting activity. Figure 4.2 represents this situation: the squares are the tweets that X posts; black ones are those which receive any retweet (circles), while white ones are not retweeted at all. X is supposed to publish a finite set P of posts (with cardinality – i.e. the total number of tweets he publishes – N_p), spread across the observation period T. In the

meanwhile other users, who are either followers or non-followers of X, are able to retweet any of those posts, if so they wish. $RT(w)$, consequently, represents the set of retweets of a single tweet w. The cardinality of this set, i.e. the cumulative number of retweets a single tweet received during T, is denoted as $N_r(w)$.
RT_W indicates the subset of P containing all the tweets of X which have been retweeted at least once, and it has cardinality N_{rtw}.

It is now possible to define in details some variables, which were already mentioned in the previous chapters, namely the retweeting frequency and the amplitude. The $RT_{frequency}$ of user X is the ratio betweet the total number of tweets which have been retweeted at least once (N_{rtw}) and the total number of tweets posted by X (N_p). This quantity actually expresses the probability of a single tweet of user X to be retweeted by another user. The $RT_{amplitude}$ of the process of user X is the average number of retweets he should expect for each of his retweeted posts. In details, it is the ratio between the sum of the number of retweets for each retweeted post ($\sum_w N_r(w)$) and the total number of tweets that have been retweeted at least once (N_{rtw}). All the formulas and variables described above are summarized in figure 4.3.

$X = user$ $\quad$ $w = a\ generic\ tweet\ of\ P$

$T = total\ time\ interval$ $\quad$ $RT(w) = set\ of\ retweets\ of\ w$

$P = set\ of\ posts\ of\ X$ $\quad$ $N_r(w) = cardinality\ of\ RT(w)$

$N_p = cardinality\ of\ P$

$RT\_W = set\ of\ tweets\ retweeted\ at\ least\ once\ (subset\ of\ P)$

$N_{rtw} = cardinality\ of\ RT\_W$

$$RT_{frequency} = \frac{N_{rtw}}{N_p} \qquad RT_{amplitude} = \frac{\sum_w N_r(w)}{N_{rtw}}$$

Figure 4.3: "in-the-large" retweeting process model formulas

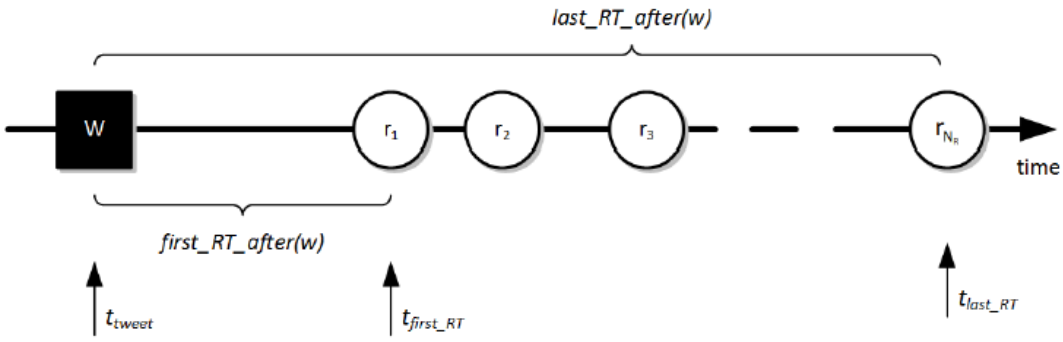

Figure 4.4: "in-the-small" retweeting process model dynamics

The retweeting process model in-the-small (figure 4.4), instead, does not take into account the user's activity; it simply focuses on each single tweet that is being retweeted, regardless who it belongs to. A

generic tweet is referred to as w and it is posted at time t_{tweet}. The set of its retweets is called $RT(w)$ and its cardinality is $N_r(w)$. Each retweet r_i of $RT(w)$, $1 \leq i \leq N_r(w)$, occurs at a precise time instant $t(r_i)$. We refer $t(r_1)$ as $t_{firstRT}$, the instant of the first retweet of the post w, while the time of the last retweet, $t(r_{N_{r(w)}})$, is called t_{lastRT}. These variables allow us to define four additional database attributes that will be used in the next chapter:

- "first_retweet_after": it is the time elapsed between the first occurrence of a retweet and the original time the tweet was posted
- "last_retweet_after": it is the time elapsed between the last occurrence of a retweet and the original time the tweet was posted
- "average_retweet_interval": it expresses the average interval of time that elapses between two retweets of the considered tweet w
- "average_retweet_time": it expresses the average interval of time that elapses before the considered tweet w receives a retweet

Figure 4.5 summarizes what stated above, expressing via mathematical formulas the attributes that have been explained through natural language.

$w = a\ generic\ tweet$

$t_{tweet} = posting\ time\ of\ w$

$RT(w) = set\ of\ retweets\ of\ w$

$N_r(w) = cardinality\ of\ RT(w)$

$r_i = i^{th}\ retweet\ of\ RT(w)$

$t(r_i) = posting\ time\ of\ tweet\ r_i$

$T_{firstRT} = t(r_1) = posting\ time\ of\ the\ first\ retweet, r_1$

$T_{lastRT} = t\left(r_{N_r(w)}\right) = posting\ time\ of\ the\ last\ retweet, r_{N_r(w)}$

$first\_retweet\_after = T_{firstRT} - t_{tweet}$

$last\_retweet\_after = T_{lastRT} - t_{tweet}$

$$avg\_retweet\_interval = \frac{T_{lastRT} - t_{tweet}}{N_r(w)}$$

$$avg\_retweet\_time = \frac{\sum_{r_i}(t(r_i) - t_{tweet})}{N_r(w)} = avg\left(t(r_i)\right) - t_{tweet}$$

Figure 4.5: "in-the-small" retweeting process model formulas

This kind of reasoning also allows the definition of the third new variable, $RT_{persistence}$, which is one of the main components of the hypotheses verification section (*refer to 5.1.1 for further details on this specific variable)*.

4.2.5 Data Summarizer

This phase operates on the table produced by the previous matcher, called "ajob" (*back in paragraph 4.2.3*), which contains one row for each pair "original tweet – retweet". Every line shows the details of both the posts (such as their respective "post_id", "text", "authors_id" and "language"), and adds the difference between their publication dates in milliseconds (new "elapsed_milliseconds" attribute).

The software gathers all the information available about each original post (quering the table with a "$GROUP\ BY\ original\_tweet\_id\ AND\ language\_id$" clause) and creates a new table, called "condensed_ajob", containing one row for every original tweet. Each of these rows contains all the information already known about an original tweet (id, language, author, publication date, text, link presence, …) and integrates it with several new database attributes:

- "number_of_retweets": the table created at the previous step ("ajob") contains one row for each pair "original_tweet – retweet" and we are grouping by "original_tweet_id", so it suffices to count the number of rows collected per group to obtain the number of retweets which matched that specific original post
- "mentions": a mention is a way to involve another Twitter user into the discussion [90], and it is performed through the use of the symbol '@', followed by the nickname of the user, without any space (e.g. "Hello @MatteoFreri how are you?") [91]. The software looks for all the '@' contained into the tweet text, then checks whether they are both preceded and not followed by a space (otherwise they could be something different, like an e-mail address, e.g. matteo.freri@polimi.it): in case a '@' passes the

test, the number of mentions contained in the tweet is increased by one. The total number of mentions computed this way is the one saved into the "mentions" field of "condensed_ajob"

- "hashtags": an hashtag is a word preceded by the symbol '#', typically used to convey the main topics of the content being posted (e.g. "I'm so #happy today! #sunshine #holidays"). Twitter gathers all the most used hashtags at the present moment and displays them on either worldwide or local charts, thus they are an extremely powerful way to generate influence, if used correctly. The process used in order to count the total number of hashtags contained into a tweet is pretty similar to the one applied to mentions: each '#' must be preceded by an empty space and directly followed by something different (e.g. "# beautifulday" will not work; though they may seem valid, hashtags like "very#beautifulday" are not considered by Twitter convention, because they are attached to another word [91]). The total number of hashtags, zero or more, is consequently saved into the table for each original tweet
- "categorized" (i.e. specificity): as stated in paragraph 4.2.1, the categorization of the tweets of the dataset follows a modified version of the Anholt model [82], and each tweet has formerly been assigned to zero or more categories by the semantic engine. Each row in the original dataset referring to the same tweet thus contains potentially different categories, with possible dissimilar sentiment. Improving the evaluation method operated by [62] [81], I counted the number of different categories a tweet was assigned to compared to the total number of rows it appeared in, instead of considering the simple binary variable that states whether it was categorized (one or more times) or not. Consequently, the value of specificity became a real number that spanned from 0 (fully uncategorized) to 1 (completely categorized). As an example, if a tweet had four rows, two with

category "Arts and culture", another one with "Exhibition" and one without any category assigned, its value of categorization (or specificity) would be

$$\frac{(1+1+1+0)}{4} = \frac{3}{4} = 0.75$$

This different approach, increasing the significance of the variable itself (that becomes real from binary), allowed the statistical analyses performed through AMOS to give much more considerable results (*see 5.1.2 and 5.2.2 for further details on this topic*)

- "length": the value of a tweet characters count is simply computed through the SQL function "$length(original\_text)$"
- "length_percentage": before starting the insertion of the rows into the condensed table, the tool calculates the average value of the lengths of all the original posts. Length_percentage is obtained dividing the length value of each original tweet by this average number. Though Twitter posts are limited to 140 chars only [92], it would be interesting to see whether the shortest ones have a greater impact on the audience than longer ones have, and this attribute gives a percentage value to make comparisons among them
- "sentiment": it is a real value, spanning from -1 to 1 (included), which expresses the feeling the author of the message wanted to convey through the tweet (-1 means completely negative, 0 neutral and 1 completely positive attitude). In the original dataset I was provided, every post was featured in many different rows, each one with a dissimilar text categorization (*refer to paragraph 4.2.1*) and a distinct sentiment value, expressed as positive (+), negative (-) or neutral (/). Those values were computed by a specific module of the same engine that classified the tweets into

categories [93]. Following a process similar to the one employed with specificity variable, I made an average of such values, converting them from chars to numbers: positive became 1, negative −1 and neutral 0. As an example, if a tweet consisted of seven different rows, two of them with positive sentiment, four with neutral and one with negative, the result would denote a slightly positive bias of the post conveyed message, being

$$\frac{[2 * (+1) + 4 * (0) + 1 * (-1)]}{7} = +\frac{1}{7}$$

In the provided dataset, 7,09% of all tweets possessed at least one sentiment classification different from neutral; such a low percentage is typical of our analysis environment, since tweets about tourism are usually descriptive or informative [61].

- "city_dimension": this attribute assumes a positive integer value, referring to the dimension of the city the tweet is about. I decided to assign to each of the cities considered in the Twitter querying process an integer, proportional to the number of inhabitants it has, starting from giving 1 to the town with the fewest people. Amalfi, with 5'000 citizens only, got a 1; Lucca and Lecce, with roughly 90'000 citizens each, got 18; Palermo (650'000 people) had 130 and Naples (just less than a million) 198; finally, Rome obtained 530, since it consists of 2'650'000 individuals. Numbers themselves start arbitrary but are carefully proportional to each other and, normalized on a small interval, serve remarkably in the hypotheses verification section as an indicator of city dimension
- "first_retweet_after": it is the time elapsed between the first occurrence of a retweet and the original time the tweet was posted (*see 5.1.1 for its employment)*
- "last_retweet_after": it is the time elapsed between the last

occurrence of a retweet and the original time the tweet was posted (*see 5.1.1 for its employment)*

- "average_retweet_interval": it expresses the average interval of time that elapses between two retweets of the considered tweet w (*see 5.1.1 for its employment)*
- "average_retweet_time": it expresses the average interval of time that elapses before the considered tweet w receives a retweet (*see 5.1.1 for its employment)*

The whole process took slightly less than two hours to complete on the available dataset, resulting in $110'714$ rows in the new "condensed_ajob" table, one for every original tweet who had at least one retweet, each possessing all the attributes specified above.

4.2.6 Retweeting Queue Evaluation

Though the considered dataset consists of hundreds of thousands of tweets, distributed over a time period of three months, it is still necessary to set a temporal threshold after which original tweets in the dataset will not be considered in the analysis. This is because the retweeting phenomenon takes some time to start, develop and come to an end, and considering into the analysis tweets which have a too little time span to be noticed and retweeted would lead to some statistical bias which can be somehow avoided. As an example, consider a tweet posted on the last day of our dataset, which received ten thousand retweets over the following week: we are not able to predict such a huge amount of retweets because our dataset ends the same day the tweet is posted, and only a small percentage of such retweets appears in our data. One possible empirical solution, which is the one adopted in this research thesis, is to consider some of the tweets in the dataset (e.g. the ones in the first

week) and check, for each of them, when their retweeting queue dampens down, i.e. understanding the approximate time instant in which most of retweets observed for that specific tweet have already taken place. The calculations performed resulted in approximately a week necessary for the retweeting queue to diminish significantly (95% of retweets were included in those $6/7$ days, average), consequently the analyses performed from now on will neglect the last week of the dataset as regards original tweets (retweets – of course – will be instead considered over the whole time span).

Here are the details of the process followed for the retweeting queue dampening time determination: firstly, the software gathers all the original tweets (pairs id-language), which belong to the first week of the dataset, from the "condensed_ajob" table. For each of these posts the tool computes the time interval between the first publication of the tweet itself and the posting time of its last retweet, determining the total number of days the retweeting queue is composed of (rounded up to the nearest greater integer). The software then establishes the number of retweets per day that the original tweet received and saves all this data in a new table, namely "analysis", which contains one row per each post_id / language / day. As an example, consider the original tweet with $post\_id = 123456$ and $language\_id = 15$, having a total of 9 retweets, 3 in the first day, 4 in the second and 2 in the third of its posting time. The new table will consequently contain three rows for the considered tweet, one for each day, with the different number of retweets just stated (*figure 4.6 shows this example*).

	analysis_id	original_tweet_id	language	day	num_retweet	week
▶	14794	123456	15	1	3	1
	14795	123456	15	2	4	1
	14796	123456	15	3	2	1

Figure 4.6: distribution of number of retweets for post 123456 language 15

After all the fetched tweets have been processed, the actual count of dampening days takes place. The test was performed at three different thresholds of retweeting queue completeness, namely 90%, 95% and 99%. The tool gets a single tweet per time and sums the number of retweets it received each day, checking after each day whether the total amount is above total_number_of_retweets*threshold. Whenever the test succeeds, it means that 90% (or 95%, or 99%, depends on which threshold the process is at the current time) of the retweets have taken place and that is the day in which the queue dampens down. An example will clarify the reasoning: imagine we have a tweet with 31 retweets the first day, 23 in the second, 46 in the third and 5 in the fourth, for a total amount of 105 retweets. At a 90% threshold, the day in which the tweet received at least $0{,}9 * 105 = 94{,}5$ retweets is the dampening one. In this case, we have a total of 54 the second day and of 100 the third, thus the dampening takes place at day three. At 95% the amount to reach is $0{,}95 * 105 = 99{,}75$, so the day is still the third. At 99%, the number goes up to $0{,}99 * 105 = 103{,}95$, so the queue requires 4 days to dampen down. Each row of a new table, "counts", will contain the id of the original tweet, its language id, the total number of retweets it got and three other attributes, each representing the number of dampening days at the three different threshold levels.

The average number of days of retweet queue dampening, computed on the whole first week of original tweets, resulted to be less than three. The tweets with fewer than six retweets represent the 95% of the whole first week (and the complete dataset itself) and the average number of days of queue dampening increases along with the number of retweets, thus I decided to check what was the average number of dampening days for those posts which had 6 retweets or more, in order to determine an approximate upper bound for it. The actual number for such a restriction ended up to be from six to seven (depending on the different considered threshold). Overall, considering the empirical nature of this test, which takes into account the first week only as representative of the whole dataset, excluding the last week of originals from the dataset seemed the wisest choice to make. This takes into account both the observations stated in this paragraph (the first number, three, gave a order of magnitude, while the second one, seven, gave the upper bound) and gives confidence that the results are as correct as possible (most of the retweeting queues should correctly dampen down in that time), since the tweets with longer queue dampening times are very rare (less than 3% overall).

This phase, overall, was quite long and took more than fourteen hours to complete on the first week of the dataset.

4.2.7 Klout Assessment

The last portion of the following section, "User data gathering and rollup", describes how the software gathers users' Klout scores, which will be used as a substitute to the number of followers on Twitter in the statistical models analysis (actually, both models will be developed in parallel). The number of Twitter followers allows to

determine whether a user can be considered an influencer or not (according to the distinction between "influence" and "influencer" explained in 2.2); the main reason I conducted this empirical Klout assessment is to establish if Klout scores could be used to determine if a user is an influencer as well.

Klout is a website launched in 2008, whose (actually misleading) slogan is "Klout: the standard for influence", which exploits social media analytic techniques to rank its users, who are individually assigned a numerical value between 1 and 100, called "Klout score" [57] [58]. As already mentioned, Klout exposes some APIs (which I queried through Klout4j) which allow the software to perform operations such as user id or Klout score retrieving. I decided to add such a score to my Twitter users in the dataset since its computation is mainly based on the user's Twitter network (size, posted content, real interactions, feedback received, …). In 2011 Sean Golliher, professor from Michigan State University, showed how the logarithm of the number of Twitter followers was enough to explain the 95% of the scores variance [105]. Over these two years Klout refined its score computation through the data collected from other social networks linked to the user account (such as Facebook, Google+, Linkedin, Foursquare, Wikipedia and Instagram). The website methodology keeps evolving and the ways scores are calculated and updated become more and more accurate, because the Klout developers team keeps adding support to new social media platforms and seems to treasure the (various) critics received for improving their own strategies; Twitter, however, remains the main source of the whole process.

The empirical evaluation performed proceeded as follows: I created by hand a list of influencers (i.e. people, brands and institutions which have a broad audience and high reputation), gathering them

from various sources, such as the "Time 100" of 2013 (which is the list of the 100 most influent people of the year) [106], the list of people who appeared in "Time" charts more than twice in history (such as Oprah Winfrey, 9 appearances, Barack Obama, 8, but also Mark Zuckerberg and Larry Page, 3, for a total amount of 18 people) [107], Esquire "Top 75 most influential people of the 21st century" of 2008 [108], Twitter 50 top people (i.e. the ones with the highest number of followers) and all the Facebook fan pages which had the most followers, in many different fields (politics, sports, music, education, news; actors, models, writers & bloggers, executives, brands), collected from a website that daily updates all the charts from Facebook, Twitter and Klout [109]. The whole research resulted in a collection of 225 different names (to make the analysis possible, only those with a Twitter account were considered). The next step was gathering Klout top score 225 users (U.S.A. football teams, which had very high scores at the time, were ignored, since their reputation was very biased by their location and they are actually of very little interest for the current analysis) from the same website used for Facebook pages and Twitter influencers [109].

After that, I developed a small application which was able to gather some information – such as number of followers and Klout score – from Twitter and Klout, about the two different sets of people (influencers and Klout tops), and stored it in two different database tables, allowing the computation of some statistics. Out of 225, 90 names were in both lists (40%), which is quite a good result considering that the criteria of selection of the 225 influencers were independent from Klout charts. Klout tops had an average score of 91,67 (standard deviation 2,95), while their average number of followers was 4′424′314 (s.d. 6′981′720,42); top influencers resulted in a score of 83,97 (s.d. 11,43) but possessed more followers average (5′115′743 average each, s.d. 7′352′127,67).

The evaluation process followed the methodology adopted by the aforementioned 2011 study [105], which considers the influence represented by the Klout score as a very simple function:

$$klout_{score} = A + B * \ln(num\_followers)$$

where A and B are to positive constants. The application cycles through various pairs of A and B among preset thresholds and determines which couple of parameters allows the highest number of accurate Klout score predictions for the given set of users. The verification was performed on the two sets, one with 225 Klout top scorers and the other with 225 top influencers, both independently and merging them.

Tables 4.1 and 4.2 summarize the results obtained letting both the parameters A and B extend through the widest possible ranges (from 0 to 99 for A and from $0{,}1$ to 145 for B, which are the threshold values that force the score not be lower than 0 or higher than 100). The Klout score is a number that varies day by day (the same Klout staff warns not to store its scores for more than five days, due to a sort of "validity expiration" [110]), so we can assume that all the results which get fewer than 5 points close to the real value are pretty accurate (table 4.1). Another threshold of 10 points (which still refers, globally, to pretty consistent results) was established to make some useful comparisons (table 4.2).

Threshold: 5	Klout top scorers	Top influencers	Both sets together
Users:	212 (94.2%)	134 (59.5%)	308 (68,4%)
Value of A:	85.5	79.3	86.2
Value of B:	0.5	0.5	0.4

Table 4.1: percentage of exactly computed Klout score with threshold = 5

Threshold: 10	Klout top scorers	Top influencers	Both sets together
Users:	225 (100%)	185 (82.2%)	410 (91.1%)
Value of A:	83.1	81.7	83.1
Value of B:	0.4	0.5	0.4

Table 4.2: percentage of exactly computed Klout score with threshold = 10

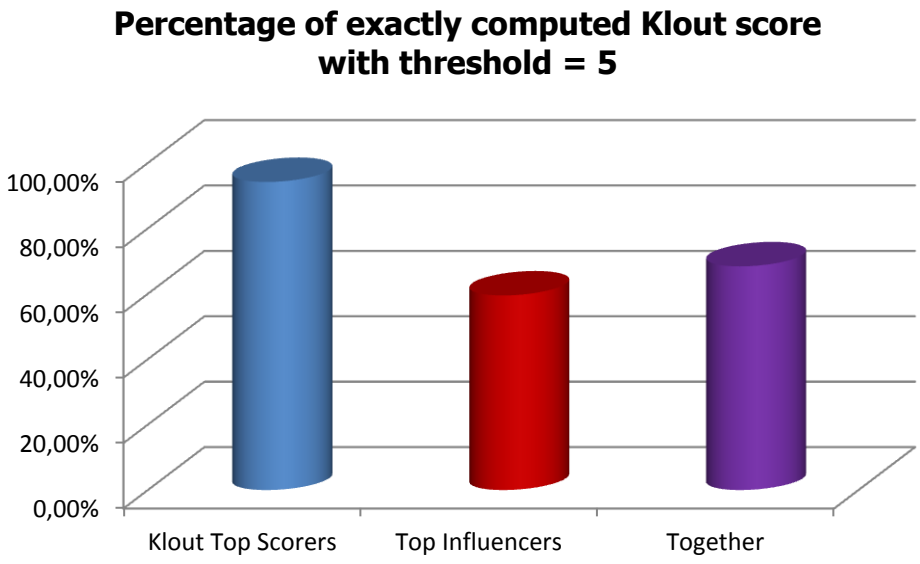

Figure 4.7: graph summarizing table 4.1

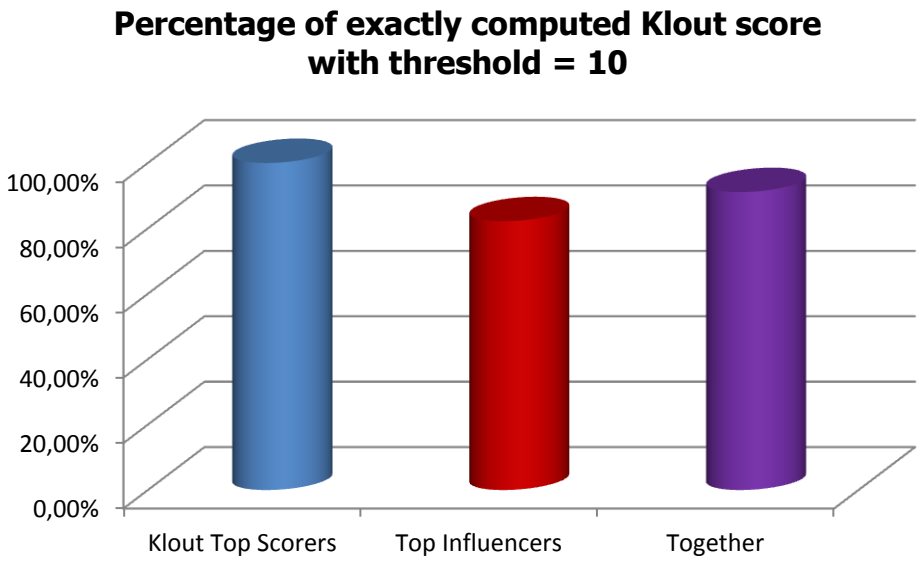

Figure 4.8: graph summarizing table 4.2

As one can easily notice, the actual Klout score value for Klout top users is really easy to predict with high accuracy (from 94.2% to 100% match with the two different thresholds), just exploiting the simple equation provided, which makes use of the number of Twitter followers only (actual values of parameters A and B are not the issue to address, though uniform among all tests). Percentages are somehow lower for top influencers (from 59.5% to 82.2%), but still the impact of that lone variable is remarkable. The analysis of the two combined sets of users gives – predictably – results which are midway the previous ones (68.4% of scores can be computed with high accuracy and 91.1% with close approximation).

A further test can get inspiration from the other equation suggested in [105], which considers not only the number of Twitter followers, but also the number of retweets a user has received. The test can be made way more significant if performed on a large dataset, that is why I decided to apply such an approach to the set of all users of my original dataset who have received at least a retweet. The total amount of those users is $64'658$ and they have an amplitude value

(*get back to paragraph 4.2.4 for further details on amplitude*), which reflects the number of retweets a user can expect to receive average for each tweet. Due to the inherent complexity of the new scenario (much bigger amount of data and one additional constant value to multiply with the number of retweets), a more structured approach was developed, using IBM AMOS Structured Equations Modeling analysis tool (*refer to 5.1.1, where it will be used for hypotheses verification*). This software allows to manage huge quantities of data and to validate the generated models, corresponding to the hypotheses made, through statistical fit indexes. The natural logarithm of the number of followers was taken in order for the variables to range within the same order of magnitude and, according to AMOS SEM guidelines, all the variables have been normalized. Standardization is a common practice to allow uniform statistical results among different units of measure [111]; linearization is useful because some dataset describing variables follow a power-law distribution [112] while SEM implies linear regression. The model is showed in the figure 4.9, while the fit indexes are show in table 4.3.

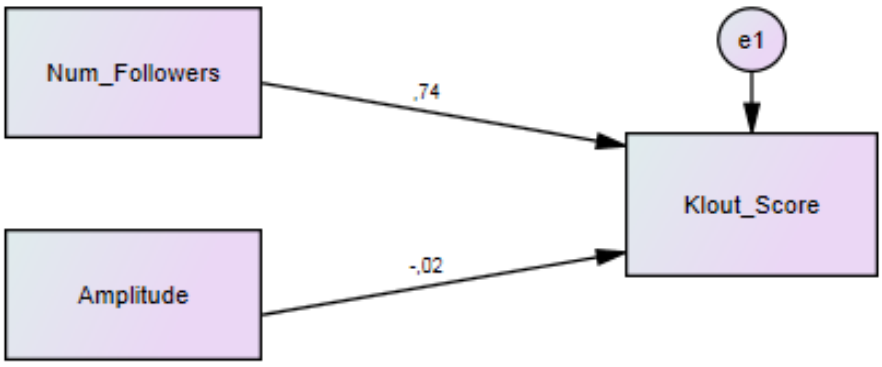

Figure 4.9: model showing Klout score heavily influenced by the number of Twitter followers

Index	Value	Desired Level	Reference
χ^2	$879{,}426$	-	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	-	-
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.98	≥ 0.95	Hu, Bentler (1999)
TLI	0.939	≥ 0.90	Tucker, Lewis (1973)
CFI	0.98	≥ 0.90	Bentler (1990)
RMSEA	0.128	-	Kenny, Kaniskan, McCoach (2013)

Table 4.3: goodness-of-fit indexes for Klout assessment model

As suggested by $[113]$, many indexes were taken into account to assess the overall fit of the produced model. The evaluation of the Chi-Square (χ^2) statistic has been proved to be particularly sensitive to sample size $[114]$ $[115]$ $[116]$: most of the times it rejects the model when large samples are used, even if it is actually supported by empirical data $[117]$. As a consequence, its value can be ignored and further indexes can be analyzed. The Normed Fit Index (NFI) compares the hypothesized model to the null model: a value very close to 1 indicates a perfect fit, thus NFI being greater than 0.95 is recommended $[118]$. The Tucker-Lewis Index (TLI) represents the improvement in the proportion of total covariance explained by the hypothesized model over that explained by the null model $[119]$, while the Comparative Fit Index (CFI) accounts for the reduction in

the model misfit of the hypothesized model compared to the null model [120]: for both those indexes, values overcoming the threshold of 0.90 are suggested [119] [120]. The proposed model perfectly satisfies all such indexes constraints, as one can check in table 4.3. The root mean square error of approximation (RMSEA) measures how well the model fit the population's covariance matrix [121]. Though a value of RMSEA lower than 0.1 is usually suggested [122], [123] pointed out how RMSEA, whose formula namely is

$$RMSEA = \frac{\sqrt{\chi^2 - d.f.}}{\sqrt{d.f.(N - 1)}}$$

, with N equal to sample size, is heavily biased in models with high N and low degrees of freedom (d.f.), thus should not be considered in model fitness checking. This is our case, since the degrees of freedom for the proposed model are minimal (equal to 1) and sample size large ($N = 64'658$, where usually 200 is enough to start a correct analysis). The model, whose correctness has just been proven, shows very little (though negative) correlation between a user amplitude and its Klout score, whereas the impact of the number of followers on that is substantial, which is what we wanted to point out from the start.

All the empirical and analytical tests performed show how the Klout score really is a parameter which can measure whether a user is an influencer or not, and thus can be safely employed in future analyses instead of the number of Twitter followers, in case that is not available for a user anymore.

4.2.8 User Data Gathering and Rollup

This last phase of my java tool is made of two steps: the first one

consists of rolling up the data of "condensed_ajob" table per user (into the new "users_condensed_ajob" relation), while the second one gathers data about Twitter users both from the social network itself and from Klout (*refer to 2.1.4 and 4.2.7 for further details on this platform*) through their APIs, enriching what is already stored in the database.

As a preliminary step, I fetched all the ("tweet_id", "language_id") pairs from the abovementioned condensed table, which are the pairs of originals with at least one retweet. Then, I added the attribute "has_retweets" to the table containing the original tweets only (such boolean attribute specifies whether an original tweet possesses any retweet or not) and marked with "true" all the rows which matched the pairs I just fetched. This is useful for the calculation of the attribute "frequency", described in the following lines.

The software is then able to start the real roll-up phase: it gets all the distinct pairs ("author_id", "language_id") from the "condensed_ajob" table, namely all the distinct authors that posted an original tweets with at least one retweet, and for each of them it computes the average of some parameters (such as "first_rtw_after", "last_rtw_after", "avg_rtw_interval", "avg_rtw_time", "categorized", "contains_link", "num_retweet" – *which actually is the user amplitude, see 4.2.5* –, "num_mentions", "num_hashtag", "length", "sentiment") from the same table. The tool also computes the author frequency attribute, namely dividing the number of retweeted tweets (which can be derived by counting the ones in the condensed table with "has_retweets" – described above – equal to one) by the total number of original tweets the author has posted (counted from the source table of original tweets). All these new attributes are saved in a new table, called "users_condensed_ajob", which contains one row for each author, with all the details just explained.

While performing such calculations, the tool can be stopped every 500 saved users, since this step is a long one (it took about eight hours in the current setup). That is a pretty arbitrary number, at which the software pauses for some seconds, allowing the user to stop the execution and to resume it later on, with no issues.

As soon as all the users data is collected and stored, the following phase is ready to begin: Twitter must be queried in order to get the additional details about each user which were not present in the dataset but are necessary for the analysis. The ones which are useful in this thesis research project are the number of followers of a user, but also other details are fetched and stored (such as nickname, provided real life name, number of people followed, dwelling place – if provided –, total amount of status ever posted, account creation date, last posted tweet, and Twitter ID) for future reference and potential applications (*refer to figures 4.10 and 4.11).*

Since the IDs in the dataset do not correspond to the ones Twitter uses, the tool has to use the usernames (also called user "screen_name") in order to distinguish among users: actually, that even makes things easier, because the username is unique all over Twitter and language distinction among users with same ID is no longer required. After collecting all the distinct author usernames from "users_condensed_ajob", the software gathers them in groups of a hundred and queries Twitter through one of its APIs, which accepts a list of up to a hundred usernames as input and returns details about each of the users in one single call. A brand new database table, called "twitter_users", contains one row per each user, with all the detailed attributes described few lines ago (*keep referring to figures 4.10 and 4.11*).

nickname	nome_cognome	following	followers	id	lingua	luogo	status	creazione	ultima_modifica	klout_score
0004JMP	Jade Marie Phillips	368	396	155590426	en	Llanelli, Wales	3669	2010-06-14 17:15:08	2013-12-22 17:44:20	27
00doppiozero	doppiozero	1287	12357	252124333	it	Milano, Italia	7615	2011-02-14 15:59:08	2013-12-22 17:28:17	61
00rocketgirl	Rocketgirl	531	2894	41480968	en	Clutch City, TX	10663	2009-05-21 01:41:25	2013-12-22 17:46:30	56
00Vintage	00vintage	288	10298	403628263	it	Roma	57781	2011-11-02 20:26:20	2013-12-22 17:21:14	42
00_Fabiana	.	2470	2192	774079158	it	Londra	8708	2012-08-22 19:02:06	2013-12-22 17:28:42	27
02Giusy	Midnight Memories ?	6062	11078	343490280	it	I need @LWTommosSmile	45418	2011-07-27 19:11:55	2013-12-22 17:29:21	38
02love_lovely	MAAD OR NAAH????	791	637	597555412	en		12770	2012-06-02 18:35:55	2013-12-22 17:31:54	32
031Marco	Marco	68	15	486014702	it		421	2012-02-07 21:41:08	2013-12-22 17:29:40	12
070Andrea	andrea manzoni	680	618	497317755	it	Milano	19846	2012-02-19 22:19:26	2013-12-22 17:20:15	34
060608it	060608	392	1873	161615595	it	Roma	1196	2010-07-01 10:50:03	2013-12-22 17:20:31	47

Figure 4.10: Sample of Twitter users database table, part 1

	tweet	klout_id
▶	Sprint technique... It's been awhile!! #hopskipjump	42221251271597419
	RT @TwitSofia_It: Una 'promessa di #felicità'.Così @Massarenti24 inizia l'intervista su @00doppiozero dedicata alla collana in collaboraz...	647893
	RT @Brooooke_H: Hawaii bound with this girl @00rocketgirl http://t.co/cmbTG9FMrG	41939776296374859
	@eliferracini poi mi spieghi come hai fatto ad entrarci NEL piumino! :)	39969452180403364
	Today stats: No new followers, 5 unfollowers via http://t.co/O7TjrHMTiv	27303082814723574
	RT @NiallOfficial: Derby v Doncaster tomorrow ! Wey hey ! @Louis_Tomlinson	32651106549489564
	RT @myBUSINESSntYOU: Mfss Get To Posting Picture Wen A Person Died & Or Lock Up , Were Was Those Pictures Wen They Was Free & Alive?	54606154191426610
	RT @mauromassimo: TROPPOFORTE Rodesia,parodiache fa' alcasodei colleghidi Melzo,quando sonoin riunionecon ildelegatopensionato http://t.co/...	74309400614575287
	Come fai a non pensare a LEI quando la ritrovi in tutte le canzoni d'amore,nella tua testa,nel tuo cuore e in tutto questo maledetto dolore.	78813001583221857
	#Musica: #Roma #Gospel Festival, dal 21 al 31 Dicembre 2013 all'Auditorium Parco della Musica: http://t.co/V11U87Ca4m	130604396810719862

Figure 4.11: Sample of Twitter users database table, part 2

Twitter, at the time the queries were performed, allowed for 180 calls (in this case, of 100 users each) every 15 minutes per application; the total number of usernames was around 65'000, making this step last about a hour. The waiting mechanism embedded into the tool makes use of a table called "limits", which contains rows with the "application_name" (Twitter, Klout, …) as primary key, and "timestamp" and "number_of_requests" as further attributes. The first time an application performs a Twitter query, its name, the timestamp of the performed action and "1" as number of requests are saved as a new row in the database. Each time a query to Twitter is needed, the software checks if 15 minutes have elapsed since the first request performed: in case they did, the number of requests is set to 0 and the timestamp is reset to the current time (i.e. a new temporal window has started), otherwise the tool checks if the query limit number (180) has been reached. If we are above such a threshold, the tool needs to wait the time necessary to complete the 15 minutes window, after which the number of requests is reset; if we are below the threshold it simply increases the number of requests performed by 1 and goes to the following query, repeating such process until all the users have their respective details stored.

The main issue to face in such an approach is that 16% of the users who posted original tweets in the dataset (which spans from December 2012 to February 2013, as already stated) were not active on Twitter anymore (due to suspension, deletion, …) at the time I tried to gather their data, so they were not saved in the "twitter_users" table. This concern would heavily reduce the size of the dataset as regards the "in-the-large" process model analysis, that is why I decided to gather users' Klout score, a somehow similar – but way more available – influence metric (*refer to 2.1.4 and 4.2.7*).

All the rows of the table which contains Twitter users details have the "klout_score" and "klout_id" attribute set to "-1", as a placeholder. Before querying Klout for users scores, the tool fetches from this table all the users who have no Klout score set yet, thus limiting the number of queries to the ones which are necessary and allowing the user to stop and resume the tool any time before phase completion.

Klout is way more strict than Twitter on API calls limits: at the time my tool queried it for data, it allowed no more than $20'000$ calls per day, at a maximum rate of 10 per second. Klout offers an API to get its id from a user Twitter id, since this first one is necessary to call the API that returns the Klout score: subsequently, two calls are required for each user in order to obtain the score, and approximately $10'000$ users can be fetched per APIs key per day. As a further optimization, once a "klout_id" has been queried it gets stored into the database, reducing the number of queries necessary for getting a user updated Klout score in the future to one only, since that is a unique identifier that does not vary in time.

While in this phase, the tool checks that 100 milliseconds have passed since the last query before proceeding to the following one, due to the Klout APIs limitations. It also checks, through the "limits" abovementioned table, that the limit of $20'000$ calls total has not been reached for the current day: in case it has, it switches to the other Klout APIs key I acquired in order to speed up the overall process. This allows to perform other $20'000$ calls in the same day, almost halving the required time. If both the keys have used all their available calls up, the tool must wait the following day in order to proceed in further querying. Klout released no documentation, at the time the tool was written, which explained how the 24 hours timeslot had to be considered: does it reset 24 hours after the first performed

call or at the beginning of a new day? In this last case, which time zone does it refer to? Is it absolute (e.g. Greenwich) or local? As I was able to determine on my own, the tool reset approximately at midnight, regardless to when the first query of the day was performed, but there is no API call to determine if the call limit was reached or not. That is the reason why I counted the number of performed requests (as I had always done) but also took into account, if the tool started to return errors (which can occur if the user does not exist or if the call limit was reached, but such situations cannot be distinguished via Klout4j library), that too many subsequent errors would mean that threshold was reached and execution had to be suspended for a while.

The number of Twitter users whose Klout score had to be retrieved were about 55'000; applying to the constraints imposed by the platform itself the gathering phase spanned over three days of time, with seven hours of active running each (three and a half for each of the application keys). After collecting such data, the tool saves them also in the "users_condensed_ajob" table previously described, if the user wishes so.

The last part of the tool considers the users who were not present on Twitter anymore (16% of the total, about 9'000), thus gathering them from the "users_condensed_table" and not from "utenti_twitter", and collects their Klout score directly into the users table. The Twitter id of the user was available in the provided dataset, so the method described before could be applied as well.

At the end of the fetch, the number of users in the whole dataset who had no Klout score was just 0.9% of the total amount, which is way lower than the percentage encountered with Twitter itself.

4.3 Conclusions

The software tool described in this chapter allows this research project to switch from the theoretical to the practical point of view. Interesting empirical results on the Klout score showed how this measure cannot actually be recognized, at least at the present moment, as representative of the user overall influence, but it can be regarded as an index indicating whether a user is an influencer (in the classical meaning) or not. Database tables named "condensed_ajob" and "users_condensed_ajob", respectively containing the data on single-post and user level, are the ones employed in the following model analyses.

The next chapter explains how the outputs of the software have been exploited with the purpose of investigating the relationship between the content of messages and social influence. This was done by means of hypotheses testing over the newly generated data.

Chapter 5 – Hypotheses Testing

5.1 User Perspective

The data sample used in the work taken as a comparison [61] is made of two different datasets, both regarding the tourism domain, consisting of tweets and retweets collected in a month period (December 2012 and January 2013, respectively). The dataset exploited in this work, instead, is one only, but consists of posts spanning over a quarter, straddling between 2012 and 2013, which includes the previous ones and is about three times the size.

5.1.1 Modeling

Tables 5.1 and 5.2 summarize the main characteristics of the considered datasets and the respective author variables. SPSS v.20 is the tool used both in this thesis work and in the previous one in order to perform the required statistical analyses [124].

Variable	New dataset	December 2012	January 2013
Number of tweets	1'511'497	427'935	529'697
Number of retweeted tweets	110'714	35'760	43'931
Number of tweeting authors	64'658	23'456	28'719
Number of retweets	415'572	104'038	131'752
Number of retweeting authors	257'242	66'227	94'415

Table 5.1: descriptive variables for each of the considered datasets

Variable (per author)	New dataset	December 2012	January 2013
Avg. number of tweets	6.47	10.53	9.60
Avg. frequency	0.71	0.57	0.59
Avg. amplitude	2.58	2.99	2.94
Avg. specificity	0.26	0.35	0.33
Avg. first_rt_after	7.04	5.20	5.85
Avg. last_rt_after	14.29	10.60	11.59
Avg. avg_rt_rime	9.35	7.11	7.95
Avg. avg_rt_interval	7.98	5.40	6.35

Table 5.2: descriptive variables per tweeting author for each of the considered datasets

The abovementioned variable $RT_{persistence}$ (*see paragraph* 4.2.4) is not directly derivable from the data, but its validity was assessed through a Principal Component Analysis (PCA) [125]. This statistical method allows to evaluate whether some metrics are enough correlated to represent coherent properties of the same phenomenon or not: in this situation, the persistence metrics we took into account are the variables "first_rtw_after", "last_rtw_after", "avg_rtw_interval" and "avg_rtw_time" (*refer to section 4.2.5 for their description*). Such a correlation was shown for these four variables regarding the two split dataset, and the analysis has been performed again on the ones corresponding to the new dataset, with the purpose of validating the aggregation of the variable $RT_{persistence}$ and proceed with the model checking on the new data. Tables 5.3, 5.4, 5.5 and 5.6 recap and compare the proposed datasets variables correlation values.

CFR	New dataset	December 2012	January 2013
first_rtw_after	0.935	0.875	0.878
last_rtw_after	0.872	0.965	0.964
avg_rtw_time	0.987	0.992	0.992
avg_time_interval	0.966	0.992	0.993

Table 5.3: composite factor reliability (CFR) of $RT_{persistence}$ variable for the three considered datasets

	first_rtw_after	last_rtw_after	avg_rtw_time	avg_time_interval
first_rtw_after	1.000	0.663	0.904	0.937
last_rtw_after	0.663	1.000	0.877	0.757
avg_rtw_time	0.904	0.877	1.000	0.930
avg_time_interval	0.937	0.757	0.930	1.000

Table 5.4: correlation matrix of $RT_{persistence}$ variable (Pearson index) for the new dataset

	first_rtw_after	last_rtw_after	avg_rtw_time	avg_time_interval
first_rtw_after	1.000	0.725	0.812	0.835
last_rtw_after	0.725	1.000	0.978	0.965
avg_rtw_time	0.812	0.987	1.000	0.988
avg_time_interval	0.835	0.965	0.988	1.000

Table 5.5: correlation matrix of $RT_{persistence}$ variable (Pearson index) for December 2012 dataset

	first_rtw_after	last_rtw_after	avg_rtw_time	avg_time_interval
first_rtw_after	1.000	0.726	0.815	0.841
last_rtw_after	0.726	1.000	0.987	0.967
avg_rtw_time	0.815	0.987	1.000	0.991
avg_time_interval	0.841	0.967	0.991	1.000

Table 5.6: correlation matrix of $RT_{persistence}$ variable (Pearson index) for Januar 2013 dataset

As one can notice, Composite Factor Reliability (CFR) values for every variable in each dataset are way over the common accepted threshold of 0.7 [126] [127], with a minimum of 0.872 in the new dataset. This, along with the high correlation factors found for all the variables pairs (the lowest value – 0.663 – still indicates a consistent correlation [128], and all others are above 0.75) and a KMO (Kaiser-Meyer-Olkin test) equal to 0.717 (above 0.5 threshold) [129], tells us that the factorization can be accepted and the four variables can be represented as a whole by the single unobserved variable $RT_{persistence}$. The eigenvalue computation for the new dataset is illustrated below: only one eigenvalue is way above 1, i.e. somehow meaningful, and it covers the 88.52% of the total variance, indicating that the resulting number of components required is one (that is, $RT_{persistence}$ only) [130] [131].

Component	**Initial eigenvalues**		
	Total	**Variance %**	**Cumulative variance %**
1	3.541	88.522	88.522
2	0.375	9.363	97.886
3	0.056	1.399	99.285
4	0.029	0.715	100.00

Table 5.7: eigenvalues and variance coverage per component for the new dataset

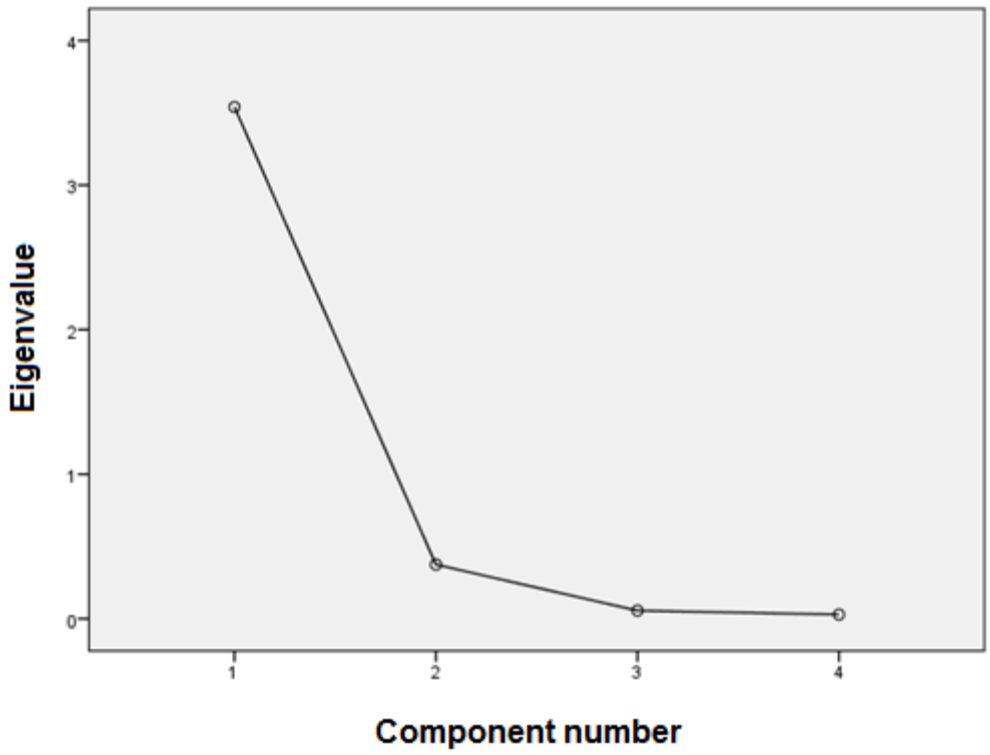

Figure 5.1: component analysis eigenvalues graph for the new dataset

The research model, briefly described in the previous sections, is here depicted. The tool employed to analyze this model is IBM AMOS version 20.0.0 [132]. It makes use of structural equation modeling (SEM), a procedure which implies second-generation data analysis techniques [133]: such methods are often brought into play in IS research, to test if a work meets the widely recognized standards for high-quality statistical analysis. While simple multivariate regression requires observed values for all variables, SEM allows the construction of models with latent (or unobserved) variables and the testing of hypotheses containing them. These variables cannot be directly measured from data and need a set of proxy variables in order to be determined: in the research model, $RT_{persistence}$ is a latent variable and it is calculated through its four proxies "first_rtw_after", "last_rtw_after", "avg_rtw_time" and

"avg_rtw_interval", that is why SEM is so valuable in this situation. According to the graphic format that SEM specifies, observed variables are expressed through rectangles, latent ones via ovals, relationships with arrows and Gaussian errors using circles attached to dependent variables. To facilitate the understanding of the diagram, the model is proposed once only, while actual numerical values are put into a summarizing table (5.8). Volumes variable is referred into the diagram as "Num_tweet".

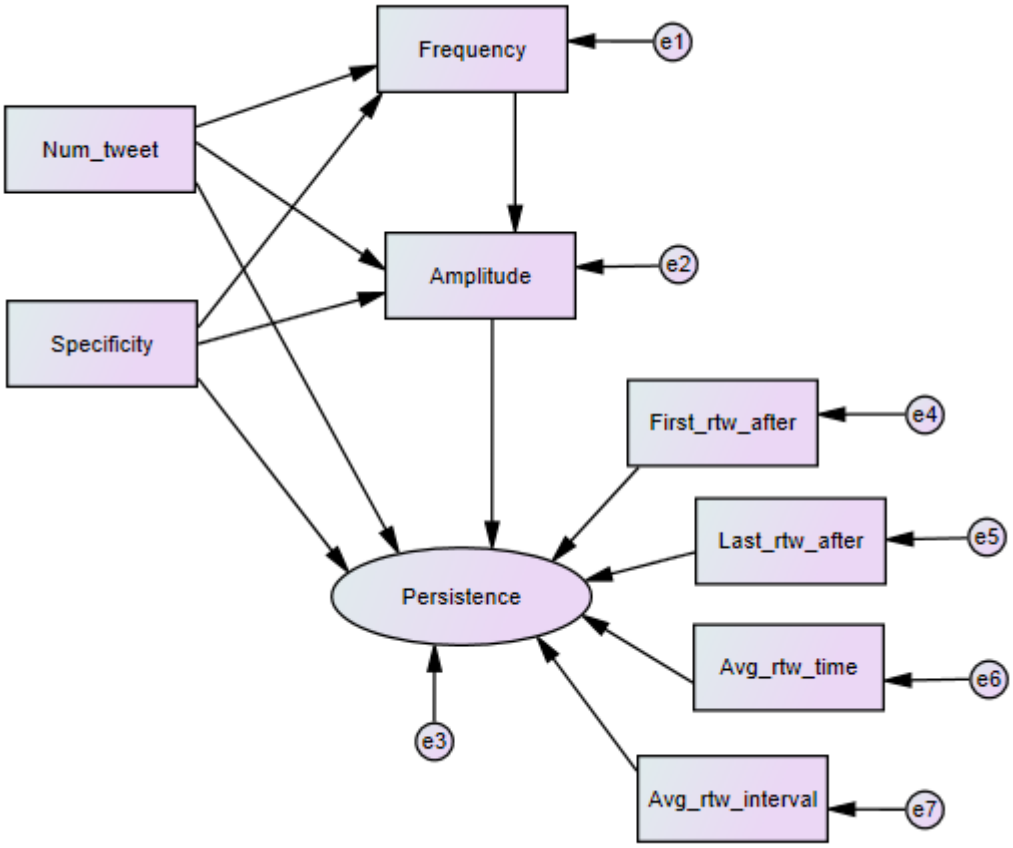

Figure 5.2: research model, user level

		New dataset			December 2012			January 2013		
Independent Variable	**Dependent Variable**	**Standardized Regression Weight**	**Standard Error**	**p-value**	**Standardized Regression Weight**	**Standard Error**	**p-value**	**Standardized Regression Weight**	**Standard Error**	**p-value**
Volumes	$RT_{frequency}$	-0.867	0.002	< 0.001	-0.928	0.002	< 0.001	-0.920	0.002	< 0.001
Volumes	$RT_{amplitude}$	0.373	0.008	< 0.001	0.375	0.017	< 0.001	0.429	0.015	< 0.001
Volumes	$RT_{persistence}$	0.085	0.004	< 0.001	-0.020	0.006	0.002	-0.033	0.006	< 0.001
Specificity	$RT_{frequency}$	0.005	0.002	0.007	0.005	0.005	0.033	0.004	0.005	0.065
Specificity	$RT_{amplitude}$	0.052	0.004	< 0.001	0.077	0.014	< 0.001	0.048	0.013	< 0.001
Specificity	$RT_{persistence}$	0.091	0.004	< 0.001	0.080	0.014	< 0.001	0.071	0.006	< 0.001
$RT_{frequency}$	$RT_{amplitude}$	0.402	0.008	< 0.001	0.367	0.017	< 0.001	0.413	0.015	< 0.001
$RT_{amplitude}$	$RT_{persistence}$	0.172	0.004	< 0.001	0.062	0.007	< 0.001	0.063	0.006	< 0.001

Table 5.8: Estimates of the regressions weights of three different datasets for the research model

Index	New dataset	December 2012	January 2013	Desired Level	Reference
χ^2	796.313	207.097	243.163	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	2	2	2	–	–
χ^2/d.f.	398.156	103.549	121.581	–	Kline (2010)
χ^2 test probability	< 0.001	< 0.001	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.992	0.996	0.996	≥ 0.95	Hu, Bentler (1999)
TLI	0.959	0.978	0.978	≥ 0.90	Tucker, Lewis (1973)
CFI	0.992	0.996	0.996	≥ 0.90	Bentler (1990)
RMSEA	0.078	0.066	0.065	≤ 0.08	Browne, Cudeck (1993)

Table 5.9: Comparison of goodness-of-fit indices of the three datasets for the research model

As suggested by [113], many fit indexes were taken into account for assessing the overall model fit (table 5.9). The evaluation of the Chi-Square (χ^2) statistic has been proved to be particularly sensitive to sample size [114] [115] [116]: most of the times it rejects the model when large samples are used, even if it is actually supported by empirical data. The relative/normed Chi-Square index (χ^2/d.f.) was proposed in [134] as an attempt to reduce such sample size weakness of the Chi-Square, but Kline, whose work we are referring to with the aim of exploiting such indexes for model assessment, discourages its use, noticing how there is "little statistical or logical foundation" to it: nonetheless, we report all the statistics into table 5.9, according to [113] and [135]. The Normed Fit Index (NFI) compares the hypothesized model to the null model: a value very close to 1 indicates a perfect fit, thus NFI being greater than 0.95 is recommended [118]. The Tucker-Lewis Index (TLI) represents the improvement in the proportion of total covariance explained by the hypothesized model over that explained by the null model [119], while the Comparative Fit Index (CFI) accounts for the reduction in the model misfit of the hypothesized model compared to the null model [120]: for both those indexes, values overcoming the threshold of 0.90 are suggested [119] [120]. The root mean square error of approximation (RMSEA) measures how well the model fit the population's covariance matrix [121], and a value lower than 0.08 is suggested by [136]. The research model satisfies all these constraints, as one can check in table 5.9 in the previous page.

5.1.2 Discussion

All three research hypotheses related to specificity (namely A, B and C) are supported by the new dataset, and that goes along to what was found in the previous research attempt. Specificity of content

plays an important role in influence generation, in terms of frequency, persistence and amplitude: on social media, it possesses a way more central position than it used to have on traditional broadcasting media, where – sometimes – the way a message is expressed can outdo the weight of the meaning itself [1]. This is true especially for Twitter, the biggest microblogging platform available these days, where the conciseness of messages forces the users to focus on the core of the information they want to convey: the testing just performed validates once again such hypotheses in that context. The variable denoting specificity assumes real values in this dataset (*refer to section 4.2.5 for further details*), while the one employed into the analysis which has been used as a comparison was binary: this allows the p-value of the relationship specificity-frequency to go below 0.01 threshold – which is regarded as the best possible for accepting an hypothesis [137] [138] –, down from the previous 0.1 value, hence reinforcing its soundness. These results, in conclusion, are a further empirical confirmation of "Attention economy" theory [139] applied to social media field, which states that a user attention is a rare good and someone who cannot find the information he is looking for quickly tries to gather it from a different source: posting authors are required, to some extent, to focus on content, in order to obtain their audience attention.

As shown in table 5.8, all empirical results support also hypotheses D, E and F. This is a further confirmation towards volumes of content being negatively associated with the retweeting frequency, but positively with retweeting amplitude: such a tendency highlights a behavioral trade-off between those two variables, indicating that public speech and social media do not actually undergo the same dynamics. Indeed, the positive effect of volumes on amplitude seems to reflect the traditional broadcasting principles, supporting the idea of a constant and heavier presence as key to marketing effectiveness

[72], but the way more negative impact on frequency strongly suggests a compromise between marketing and social media attitude for obtaining the best possible outcome: a brand should in principle concentrate on content specificity in order to attract a solid base of customers, then it could balance that with volumes in order to achieve a greater audience.

Hypothesis F, which binds volumes of content to retweets persistence, has been shown as fully supported: in contrast to what was obtained in the comparison analysis, we can state how increasing volumes of posting can actually help reinforcing a brand presence, strengthening the persistence of its tweets. This result is supported both by the fact that the regression value found is three or four times higher (0.085 in contrast with -0.020 and -0.033), that this dataset is way bigger than the previous one (since it includes both the other two, with the addition of a month period) and that the new obtained p-value is better. Anyway, further analyses on this relationship are encouraged, especially on a different dataset or even on a different field, in order to assess the validity of this hypothesis on a more general scope.

5.1.3 Control Variables

The state of "influencer" that a user possesses is, however, something that cannot be neglected: this is the reason why adding a control variable – such as the number of Twitter followers or the Klout score (*refer to paragraph 4.2.7 for an assessment of Klout score as a metric of influencer status*) – to the model allows us to explore retweeting dynamics on a broader perspective, which includes both major and minor impact variables.

This analysis, which extends the work described in the previous paragraph, was performed on the new dataset only, and the statistical tools used were still SPSS and AMOS, as previously stated. The number of Twitter users of whom I possessed the number of followers was way lower than the total number of available users who had posted at least a retweeted tweet ($54'587$ out of $64'658$ total, 16% less); this is because many of those accounts were not available on Twitter anymore at the time of the number of followers fetch, so they could not be retrieved. That is why I used the Klout score of a user as a second control variable: it allowed the majority of the users to be taken into account in the model ($63'909$ out of $64'658$, almost 99%), making the dataset statistically more suitable and comparable to the one used in the previous analyses (*refer to 5.1.1*). The research model implied is the same as before, but control variables were added once per time, in order to test the impact of volumes and specificity when a user reputation is considered. As usual, models are depicted without values (figure 5.3 and 5.4), which are instead reported into separated tables (5.10 and 5.12).

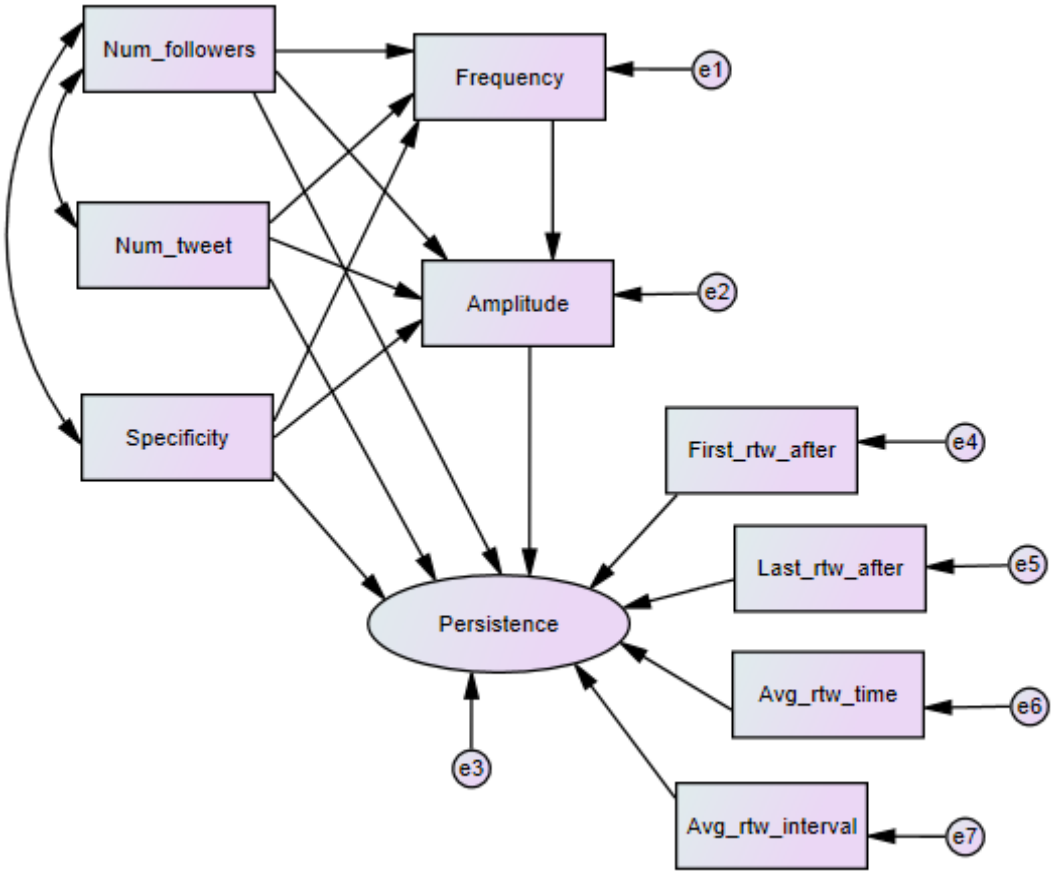

Figure 5.3: research model with control variable "Number of Twitter Followers"

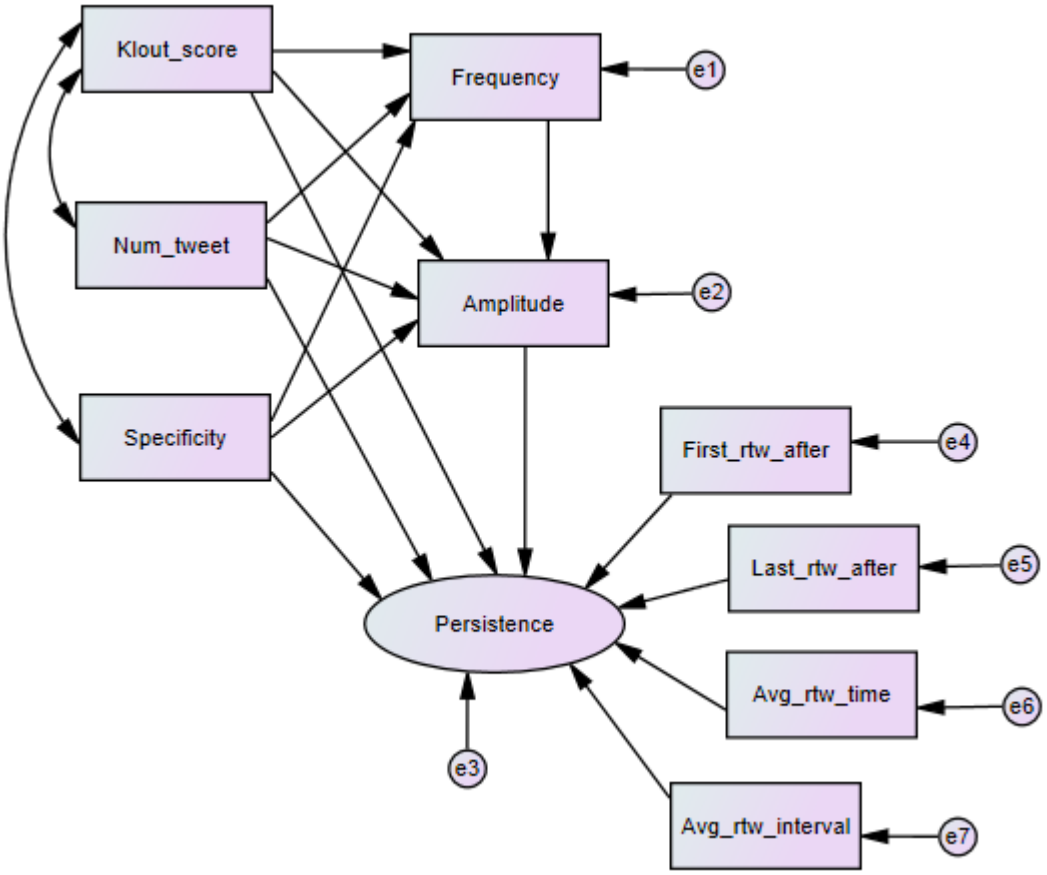

Figure 5.4: research model with control variable "Klout Score"

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Volumes	$RT_{frequency}$	-0.852	0.002	< 0.001
Volumes	$RT_{amplitude}$	0.366	0.008	< 0.001
Volumes	$RT_{persistence}$	0.094	0.004	< 0.001
Specificity	$RT_{frequency}$	0.004	0.002	0.038
Specificity	$RT_{amplitude}$	0.048	0.004	< 0.001
Specificity	$RT_{persistence}$	0.084	0.004	< 0.001
Num_followers	$RT_{frequency}$	0.029	0.002	< 0.001
Num_followers	$RT_{amplitude}$	0.234	0.004	< 0.001
Num_followers	$RT_{persistence}$	-0.008	0.004	0.048
$RT_{frequency}$	$RT_{amplitude}$	0.398	0.008	< 0.001
$RT_{amplitude}$	$RT_{persistence}$	0.165	0.004	< 0.001

Table 5.10: Estimates of the regressions weights for the research model with control variable "Number of Twitter followers"

Index	New dataset	Desired Level	Reference
χ^2	654.059	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	2	–	–
χ^2/d.f.	327.030	–	Kline (2010)
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.992	≥ 0.95	Hu, Bentler (1999)
TLI	0.940	≥ 0.90	Tucker, Lewis (1973)
CFI	0.992	≥ 0.90	Bentler (1990)
RMSEA	0.078	≤ 0.08	Browne, Cudeck (1993)

Table 5.11: Goodness-of-fit indices for the research model with control variable "Number of Twitter followers"

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Volumes	$RT_{frequency}$	-0.873	0.002	< 0.001
Volumes	$RT_{amplitude}$	0.265	0.008	< 0.001
Volumes	$RT_{persistence}$	0.086	0.004	< 0.001
Specificity	$RT_{frequency}$	-0.001	0.002	0.624
Specificity	$RT_{amplitude}$	0.032	0.004	< 0.001
Specificity	$RT_{persistence}$	0.093	0.004	< 0.001
Klout_score	$RT_{frequency}$	0.082	0.002	< 0.001
Klout_score	$RT_{amplitude}$	0.294	0.004	< 0.001
Klout_score	$RT_{persistence}$	-0.034	0.004	< 0.001
$RT_{frequency}$	$RT_{amplitude}$	0.305	0.007	< 0.001
$RT_{amplitude}$	$RT_{persistence}$	0.182	0.004	< 0.001

Table 5.12: Estimates of the regressions weights for the research model with control variable "Klout score"

Index	New dataset	Desired Level	Reference
χ^2	802.047	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	2	–	–
χ^2/d.f.	421.024	–	Kline (2010)
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.992	≥ 0.95	Hu, Bentler (1999)
TLI	0.939	≥ 0.90	Tucker, Lewis (1973)
CFI	0.992	≥ 0.90	Bentler (1990)
RMSEA	0.08	≤ 0.08	Browne, Cudeck (1993)

Table 5.13: Goodness-of-fit indices for the research model with control variable "Klout score"

As one can check in the fit indexes table reported (5.11 and 5.13), the two different datasets suit the models for their respective control variable scenario (*see 5.1.1 for a detailed description of the indexes*).

As regards the impact of that control variable on the parameters describing the retweeting dynamics, higher values can be observed in the results obtained through Klout score employment: that variation is due to the closer connection that such a number possesses with variables like frequency, persistence and amplitude, because Klout adjusts its scores taking also those parameters into account, even if slightly. Regardless of pure numerical values, both models imply a strong positive relationship between number of followers (or Klout score) and amplitude: this highlights how being an influencer is still a key factor in retweeting dynamics, because it automatically implies a wider audience and generates a broader amount of retweets, giving little importance to other variables.

Still, both set of results validate what was stated in the previous pages regarding specificity (except relationship specificity – frequency, whose p-value is either quite or very high and does not allow statistical deduction): no matter how important you are, talking about a definite subject still has an impact, though small if compared to some other effects, on how your audience reacts. Volumes were confirmed to be very negatively related to retweeting frequency and positively related to amplitude, in both cases. This time as well, data seem to validate a positive correlation between volumes and persistence, as a further support to what was found in the previous hypotheses tests (*back in 5.1.2*).

In conclusion, the importance of being an influencer cannot be neglected, but also other variables – such as specificity of content – have an impact on the readers perception and must be taken into

account both by normal users and established companies for achieving the best available result.

As regards corporations, a successful online marketing strategy – i.e. able to attract potential customers and to keep the attention of those who already are – cannot be based on volumes only, but should rely on specificity of content too (as already stated for public speaking [66]): these two factors, combined, help managing the trade-off which exists between retweeting frequency and amplitude (*refer to paragraph 5.1.2 where this conclusion was made*). Given the undeniable importance of an established amount of active readers, advertising approaches ought to concentrate on gaining a wider audience first, then on ensuring its constant attention.

5.2 Single-Post Perspective

For this model testing phase, the set of tweets which were retweeted at least once is considered, otherwise they would possess no metrics for influence evaluation, like amplitude (*refer to section 3.2*).

5.2.1 Modeling

The research model applied in this context is depicted in figure 5.5. As usual, for the sake of clarity, actual numerical values are stored into separated tables. The idea underlying this new research model consists in taking the one used at user level and constraining it with the limitations exposed in the hypotheses section (*see the beginning of 3.2*): the main aim is gaining a better understanding of the impact of different behavioral variables (such as sentiment) over retweeting dynamics, especially in the single-post perspective, which does not

take the importance of the user into account. Specificity of content and post sentiment, independent variables, are related to the number of retweets received and persistence of retweeting, which are regarded as dependent metrics of influence.

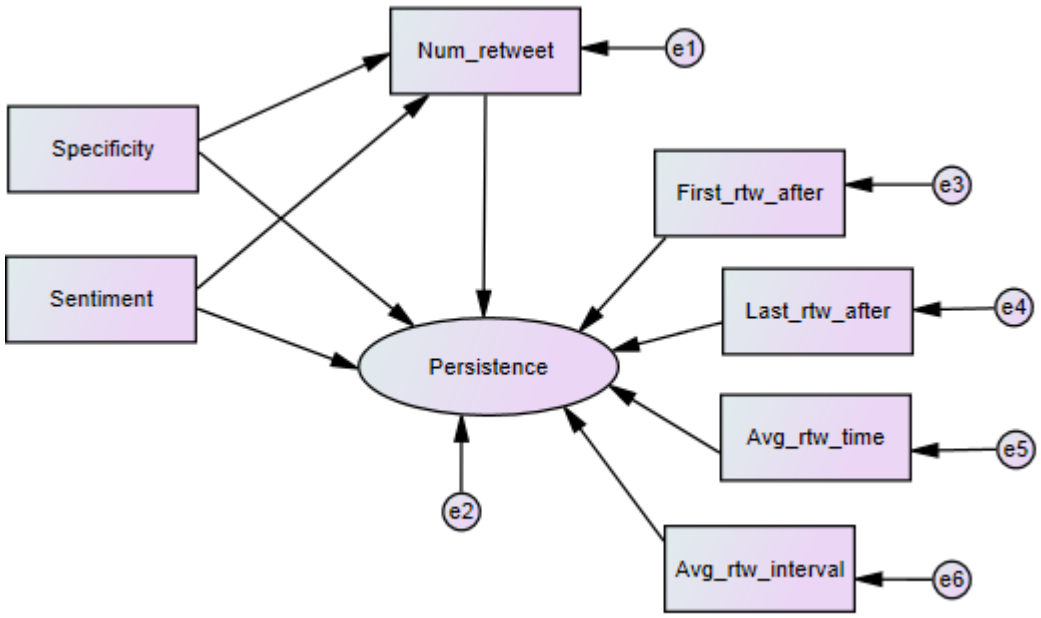

Figure 5.5: research model, single post level

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Specificity	Num_retweet	0.033	0.003	< 0.001
Specificity	$RT_{persistence}$	0.076	0.003	< 0.001
Sentiment	Num_retweet	-0.017	0.003	< 0.001
Sentiment	$RT_{persistence}$	0.012	0.003	< 0.001
Num_retweet	$RT_{persistence}$	0.196	0.003	< 0.001

Table 5.14: Estimates of the regressions weights for the research model at single-post level

Index	New dataset	Desired Level	Reference
χ^2	0.384	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	–	–
χ^2 test probability	0.536	> 0.05	Bollen, Long (1993)
NFI	1.000	≥ 0.95	Hu, Bentler (1999)
TLI	1.001	≥ 0.90	Tucker, Lewis (1973)
CFI	1.000	≥ 0.90	Bentler (1990)
RMSEA	0.000	≤ 0.08	Browne, Cudeck (1993)

Table 5.15: Goodness-of-fit indices for the research model at single-post level

Referring to the fit indexes reported on table 5.15, we can see that the model has a perfect fit for the provided data. NFI, TLI and CFI are all equal to 1, RMSEA is exactly 0 and the χ^2 test has proven successful. That is very impressive, given the dataset dimension ($N = 100'218$), because usually such a fit is way more common when small sample sizes (i.e. usually around some hundreds) are involved $[117]$ $[140]$.

$5.2.2$ Discussion

Hypotheses regarding specificity, namely G and H, were both confirmed. As stated in the previous sections, such a variable has a overall positive effect on influence (table 5.14): in this specific case, at single-post level, this translates into relationships with positive values over the number of retweets received and the $RT_{persistence}$. Even when the user is not taken into account, it is highlighted how being focused on a topic while writing is a key factor for gaining attention, both for the everyday person and the established firm.

The first hypothesis on sentiment, that is I, was proven too. The feeling conveyed with what is being posted is negatively associated with the number of retweets received, meaning that tweets which somehow carry negative sentiment attract more attention. We can obtain this conclusion by observing that the sentiment variable employed can assume negative values only if the majority of the features in the tweet carry a negative sentiment, and the -0.017 value on the relationship implies that posts carrying mostly negative sentiment usually generate more retweets. Consequently, such a tendency is being shown as a common feature between social and traditional media (*refer to section 3.2 for their comparison*).

Hypothesis J, on the other hand, is not supported by empirical data: though on a small extent, a positive correlation between sentiment and persistence of content is underlined.

Supporting and extending the guidelines provided up to now (*back in 5.1.2 and 5.1.3*), a user who wants to increase its influence in the social media field should initially concentrate on posting content specifically referred to a subject. Without forgetting the positive bias of self-promotion that pervades the social media environment [74] [75], he should share messages carrying also negative sentiment as well, in order to gain a stable and focused audience. That does not mean spamming gloomy messages around Twitter: in the tourism field, it could imply writing tweets which describe a common unlucky situation the readers can identify themselves in (e.g. you state you cannot find any parking spot available in a big city) and, if you are a company, posting another tweet to propose an innovative solution (e.g. you are the owner of a startup which sells an app that finds nearest free parking spots). An alternative could be tweeting about current news, especially if referring to unfortunate and wide-appealing events. After a solid, though not massive, user base has been acquired, the user can move to increasing the number of posted tweets and balance more between negative and positive sentiment messages, with the aim of increasing the persistence of the retweeting activity and expanding his audience.

5.2.3 Clustering

This final section describes the results obtained by dividing the original dataset into clusters, which have been tested on the same single-post model (*refer to figure 5.5 from now on*): though many different classifications were carried out, only those whose model

fitness was good enough to infer some statistical deductions are reported. These results are pretty domain-specific, and should not be regarded as general-purpose guidelines, because they may differ from what previously stated due to close relationship with the tourism field.

5.2.3.1 Language Clustering

The first clustering split the original dataset by language: since the languages taken into account were two, namely English and Italian, two clusters were obtained and tested. Out of the 110'218 tweets contained in the whole set of data, 54'586 were in English and 45'632 in Italian. Tables 5.16, 5.17, 5.18 and 5.19 summarize fit indexes and standardized regression weights for both the clusters. Both models show good fitness to data, though TLI is not ideal for any of them (especially for the Italian tweets cluster, see tables 5.17 and 5.19).

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Specificity	Num_retweet	0.066	0.004	< 0.001
Specificity	$RT_{persistence}$	0.108	0.004	< 0.001
Sentiment	Num_retweet	-0.022	0.004	< 0.001
Sentiment	$RT_{persistence}$	0.013	0.004	0.002
Num_retweet	$RT_{persistence}$	0.187	0.004	< 0.001

Table 5.16: Estimates of the regressions weights for the English language cluster

Index	New dataset	Desired Level	Reference
χ^2	65.379	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	–	–
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.979	≥ 0.95	Hu, Bentler (1999)
TLI	0.875	≥ 0.90	Tucker, Lewis (1973)
CFI	0.979	≥ 0.90	Bentler (1990)
RMSEA	0.034	≤ 0.08	Browne, Cudeck (1993)

Table 5.17: Goodness-of-fit indices for the English language cluster

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Specificity	Num_retweet	-0.012	0.005	0.014
Specificity	$RT_{persistence}$	0.037	0.005	< 0.001
Sentiment	Num_retweet	-0.008	0.005	0.087
Sentiment	$RT_{persistence}$	0.006	0.005	0.208
Num_retweet	$RT_{persistence}$	0.206	0.005	< 0.001

Table 5.18: Estimates of the regressions weights for the Italian language cluster

Index	New dataset	Desired Level	Reference
χ^2	68.412	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	–	–
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.967	≥ 0.95	Hu, Bentler (1999)
TLI	0.807	≥ 0.90	Tucker, Lewis (1973)
CFI	0.968	≥ 0.90	Bentler (1990)
RMSEA	0.038	≤ 0.08	Browne, Cudeck (1993)

Table 5.19: Goodness-of-fit indices for the Italian language cluster

All relationships involved into the English tweets cluster have an optimal p-value and confirm what was formerly underlined (see tables 5.16 and 5.18). Specificity possesses a positive correlation with both influence variables and coefficients are higher than the ones found in the general-purpose model: that suggests a greater importance for foreign tourists in posting specific content, in order to attract their audience attention, because they are someway seen as independent and trustworthy source of information about the cities they are visiting. Sentiment is still negatively related to the number of retweets received and slightly positively with persistence, implying dynamics similar to those already depicted before (*in 5.2.2*). As regards the Italian cluster, due to high p-values, only relationships between specificity and influence can lead to noteworthy inferences: there is still a positive correlation between specificity and persistence, while a negative – though very small – connection is shown with the number of retweets. That differs from general guidelines and implies that readers do not usually trust the opinion of people who come from the same country as the cities they are going to visit, maybe because they could be biased by their feelings for their nation or somehow self-promoting for marketing reasons.

5.2.3.2 City Dimension Clustering

Another clustering activity was performed on the dataset, considering the dimension of the cities which the gathered tweets are about (*see paragraph 4.2.1 for a detailed listing*): two sets were identified, the former containing all the tweets regarding cities with more than one million inhabitants (only Rome in this case, for a total amount of 69'081 tweets) and the latter with the ones with fewer dwellers (every other town, 31'137); numerical results are shown

into respective tables (5.20 and 5.22). This time, while TLI is still not optimal for smaller cities, the fitness is almost perfect for data referring to bigger ones (tables 5.21 and 5.23).

In big cities context, all relationships between behavioral variables and influence carry the same sign as the ones identified in the general guidelines. The main difference is that the relationship connecting specificity and number of retweets is stronger than before (the weight is doubled), representing the need of not being too widespread: when there are many things to talk about – which is quite a common situation in a big city such as Rome – the audience appreciates the ability to go straight to the point, without generalizing too much. Such a tendency is reverted in smaller towns: that underlines how being general is preferred in these circumstances, where there are fewer chances of being both detailed and useful for a large part of the readers. When the message is precise already, due to the context, specificity impact stops playing a central role in communication.

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Specificity	Num_retweet	0.066	0.004	< 0.001
Specificity	$RT_{persistence}$	0.092	0.004	< 0.001
Sentiment	Num_retweet	-0.021	0.004	< 0.001
Sentiment	$RT_{persistence}$	0.010	0.004	0.009
Num_retweet	$RT_{persistence}$	0.198	0.004	< 0.001

Table 5.20: Estimates of the regressions weights for the big cities cluster

Index	New dataset	Desired Level	Reference
χ^2	10.781	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	–	–
χ^2 test probability	0.001	> 0.05	Bollen, Long (1993)
NFI	0.997	≥ 0.95	Hu, Bentler (1999)
TLI	0.985	≥ 0.90	Tucker, Lewis (1973)
CFI	0.997	≥ 0.90	Bentler (1990)
RMSEA	0.012	≤ 0.08	Browne, Cudeck (1993)

Table 5.21: Goodness-of-fit indices for the big cities cluster

Independent Variable	Dependent Variable	Standardized Regression Weight	Standard Error	p-value
Specificity	Num_retweet	-0.036	0.006	< 0.001
Specificity	$RT_{persistence}$	0.041	0.006	< 0.001
Sentiment	Num_retweet	-0.011	0.006	0.060
Sentiment	$RT_{persistence}$	0.014	0.006	0.010
Num_retweet	$RT_{persistence}$	0.186	0.006	< 0.001

Table 5.22: Estimates of the regressions weights for the small cities cluster

Index	New dataset	Desired Level	Reference
χ^2	25.804	–	Bentler, Bonett (1980) & Jöreskog, Sörbom (1993)
d.f.	1	–	–
χ^2 test probability	< 0.001	> 0.05	Bollen, Long (1993)
NFI	0.979	≥ 0.95	Hu, Bentler (1999)
TLI	0.876	≥ 0.90	Tucker, Lewis (1973)
CFI	0.979	≥ 0.90	Bentler (1990)
RMSEA	0.028	≤ 0.08	Browne, Cudeck (1993)

Table 5.23: Goodness-of-fit indices for the small cities cluster

Chapter 6 – Conclusions

This thesis work, after describing social media information distribution dynamics through mathematical models, provides users with tactical guidelines suitable for increasing their influence. These principles should not be regarded as either dogmatic or comprehensive: they are general conclusions which ought to be used as the basis for a more complex marketing strategy. What was obtained is evidence that content is fundamental for increasing the spread of posts on social media platforms, both for influencers and common users.

First hypotheses and subsequent model checking are about retweeting dynamics at user perspective: specificity is proven to possess a positive effect on all variables describing influence (frequency, amplitude and persistence), while volumes are shown to possess a strongly negative correlation with frequency but positive relationships with the remaining variables. A trade-off between frequency and amplitude, as regards volumes, is highlighted, illustrating how a compromise between traditional broadcasting principles and social media marketing strategies is a key factor for success. Specificity is shown to keep a role in influence generation even while user popularity (measured through two distinct variables, namely the number of Twitter followers and the user Klout score) is taken into account. An empirical assessment of Klout score is presented, showing how it can be computed via linear interpolation of the number of Twitter followers. This process makes the Klout score a suitable measure of the status of influencer of an author and

it resizes the importance of that score as an index of the user overall influence.

Afterwards, the focus is moved to the single-post point-of-view: the difference between influence and influencers is here stressed. The level of detail of the message content, i.e. specificity, is still proven to possess a positive impact on the generated influence, even when the original posting user is completely ignored. The empirical findings about the other behavioral variable considered in this model, namely the tweet sentiment, reveal that messages carrying negative feelings are more retweeted than those which possess a positive bias: such a tendency is common both to social media and traditional broadcasting means. A positive – though small – correlation holds instead between sentiment and message persistence: always keeping in mind the importance of specificity, a user should initially concentrate on acquiring a firm audience by posting messages with some sort of negativity, then a balance between distinct feelings is suggested, with the aim of arousing the user base reaction and assessing messages persistence and retweets to a higher level. All presented results can be considered pretty solid, since the dataset shows a perfect match to the employed model (all fit indexes have optimal values). Finally, two different clusters analyses are performed, namely on post languages and dimension of the cities involved in the tweets, for understanding at which point specificity of content stops playing a role in influence generation. Results can be regarded as domain-related and show how specificity is way more important while talking about broader-ranging subjects (such as the ones related to big cities) or when the author is a tourist: his opinion is considered much more valuable and trustworthy, due to his impartiality towards the place he is visiting.

As all research works do, this one shows some limitations as well. A

limit of the adopted approach is that it makes use of data concerning the tourism domain only (although complex and multifaceted), and such a thing could constraint the validity of some of the assertions to a narrower perspective. Even if wide-ranging analyses have been performed, in order to overcome the intrinsic complexity of these phenomena – which involve human-related variables –, further research would be useful to extend empirical verifications to other domains. In addition, the considered dataset covers a limited time range (three months), so the empirical conclusions may not be representative of the whole phenomenon. Future work should also consider more extended time frames.

An interesting addition, which is left to future related works, would be adapting and validating both these models and guidelines across different social networks. Finally, taking into account different behavioral variables in the models (such as sentiment at user level, or the number of mentions in the tweet at both viewpoints), albeit it may require different approaches to the problem, could lead to a more wide-ranging understanding of influence generation and retweeting dynamics.

Bibliography

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, G. Mishne, "Finding high-quality content in social media", *ACM*, 2008

[2] L. Cantoni, A. Mandelli, E. Marchiori, "Tourists and destination management organizations facing social media and eword-of-mouth. A research in Italy", *The UCLA Anderson Business and Information Technologies (BIT) Project: A Global Study of Business Practice*, page 225, 2012

[3] C. Dellarocas "The digitization of word of mouth: promise and challenges of online feedback mechanisms", *Management Science*, Volume 49, Issue 10, pages 1407–1424, 2003

[4] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, "Everyone's an influencer: quantifying influence on Twitter", *4th ACM international conference on Web search and data mining*, pages 65–74, 2011

[5] W. J. Carl, "Current thinking on research and measurement of Word of Mouth marketing", *WOMMA (WOM Marketing Association)*, 2007

[6] M. Richins, "Negative word-of-mouth by dissatisfied consumers: a pilot study", *The Journal of Marketing*, Volume 47, Issue 1, pages 68–78, 1983

[7] D. Watts, "Influence and attention on Twitter", *Microsoft*

Research, 2013

[8] L. C. Freeman, "Centrality in social networks conceptual clarification", *Social networks*, Volume 1, Issue 3, pages 215–239, 1979

[9] B. Choi, T. S. Raghu, A. Vinze, "An empirical study of standards development for e-businesses: A social network perspective", *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, Volume 6, page 139

[10] R. T. Sparrowe, R. C. Liden, S. J. Wayne, M. L. Kraimer, "Social networks and the performance of individuals and groups", *Academy of management journal*, Volume 44, Issue 2, pages 316–325, 2001

[11] L. Hossain, A. Wu, K. Chung, "Actor centrality correlates to project
based coordination", *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* , pages 363–372, 2006

[12] S. Wasserman, K. Faust, "Social network analysis: methods and applications", *Cambridge university Press*, Volume 8, 1994

[13] S. Brin, L. Page, "The anatomy of a large-scale hypertextual web search engine", *Computer networks and ISDN systems*, Volume 30, Issues 1 – 7, pages 107–117, 1998

[14] S. Asur, W. Galuba, B. Huberman, D. Romero, "Influence and passivity in social media", *Machine Learning and Knowledge Discovery in Databases*, pages 18–33, 2011

[15] H. Kwak, C. Lee, H. Park, S. Moon, "What is twitter, a social network or a news media?", *19th international conference on World Wide Web (ACM)*, pages 591–600, 2010

[16] C. Lee, H. Kwak, H. Park, S. Moon, "Finding influentials based on the temporal order of information adoption in twitter", *Proceedings of the 19th international conference on World wide web (ACM)*, pages 1137–1138, 2010

[17] A. Leavitt, E. Burchard, D. Fisher, S. Gilbert, "The influentials: New approaches for analyzing influence on twitter", *Webecology Project*, 2009

[18] C. D'Adda, "Social media intelligence: l'analisi della influence nel microblogging", *Master of Science degree thesis,* 2010

[19] F. Benevenuto, M. Cha, K. P. Gummadi, H. Haddadi, "Measuring user influence in twitter: The million follower fallacy", *International AAAI Conference on Weblogs and Social (ICWSM10)*, pages 10 – 17, 2010

[20] D. Zhao, M. B. Rosson, "How and why people twitter: the role that micro-blogging plays in informal communication at work", *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252, 2009

[21] M. Naaman, J. Boase, C. H. Lai, "Is it really about me?: message content in social awareness streams", *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192, 2010

[22] B. Suh, L. Hong, P. Pirolli, E. H. Chi, "Want to be retweeted?

Large scale analytics on factors impacting retweet in twitter network", *IEEE*, pages 177– 184, 2010

[23] Y. Li, Y. Shiu, "A diffusion mechanism for social advertising over microblogs", *Decision Support Systems*, 2012

[24] L. Dey, B. Gaonkar, "Discovering regular and consistent behavioral patterns in topical tweeting", *21st International Conference on Pattern Recognition (ICPR)*, pages 3464 – 3467, 2012

[25] A. Tejeda-Gomez, M Sànchez-Marrè, J. M. Pujol, "TweetStimuli: discovering social structure of influence", *International Journal of Complex Systems in Science*, Volume 2, Issue 1, pages 33 – 36, 2012

[26] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs", *USENIX Association*, page 3, 2010

[27] S. Petrovic, M. Osborne, V. Lavrenko, "Rt to win! Predicting message propagation in twitter", *Prof. of AAAI on Weblogs and Social Media*, 2011

[28] J. Li, W. Peng, T. Li, T. Sun, "Social network user influence dynamics prediction", *Web Technologies and Applications*, pages 310– 322, 2013

[29] L. Bruni, C. Francalanci, P. Giacomazzi, "The role of multimedia content in determining the virality of social media information", *Information*, Volume 3, Issue 3, pages 278– 289, 2012

[30] C. Xiao, Y. Zhang, X. Zeng, Y. Wu, "Predicting user influence in

social media", *Journal of Networks*, Volume 8, Number 11, 2013

[31] A. Bruns, S. Stieglitz, "Towards more systematic Twitter analysis: metrics for tweeting activities", *International Journal of Social Research Methodology*, Volume 16, Issue 2, pages 91 – 108, 2013

[32] D. Laniado, P. Mika, "Making sense of twitter", *The Semantic Web –ISWC*, pages 470– 485, 2010

[33] O. Tsur, A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities", *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643– 652, 2012

[34] M. Efron, "Hashtag retrieval in a microblogging environment", *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787– 788, 2010

[35] F. Abel, Q. Gao, G. J. Houben, K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web", *The Semanic Web: Research and Applications*, pages 375– 389, 2011

[36] E. Gjeci, R. Gonzalez Gomarasca, "Modellazione del comportamento sociale come funzione del contenuto dei post", *Master of Science degree thesis,* 2012

[37] M. Nagarajan, H. Purohit, A. Sheth, "A qualitative examination of topical tweet and retweet practices", *Fourth International AAAI Conference on Weblogs and Social Media*, 2010

[38] A. Aue, M. Gamon, "Customizing sentiment classifiers to new domains: a case study", *Proceedings of RANLP*, 2005

[39] J. Blitzer, M. Dredze, F. Pereira, "Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification", *Proceedings of the 45th Annual Meeting - Association For Computational Linguistics*, pages 432–439, 2007

[40] E. Breck, Y. Choi, C. Cardie, "Identifying expressions of opinion in context", *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2683–2688, 2007

[41] S. R. Das, M. Y. Chen, "Yahoo! for Amazon: sentiment extraction from small talk on the web", *Management Science*, Volume 53, Issue 9, pages 1375–1388, 2007

[42] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis", *Proceedings of the 20th international conference on Computational Linguistics (ACL)*, page 841, 2004

[43] N. Godbole, M. Srinivasaiah, S. Skiena, "Large-scale sentiment analysis for news and blogs", *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 219–222, 2007

[44] A. Kennedy, D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters", *Computational Intelligence*, Volume 22, Issue 2, pages 110–125, 2006

[45] T. Nasukawa, J. Yi, "Sentiment analysis: Capturing favorability

using natural language processing", *Proceedings of the 2nd international conference on Knowledge capture (ACM)*, page 77, 2003

[46] M. Thomas, B. Pang, L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts", *Proceedings of the 2006 conference on empirical methods in natural language processing (ACL)*, pages 327–335, 2006

[47] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, "Learning subjective language", *Computational Linguistics*, Volume 30, Issue 3, pages 277–308, 2004

[48] E. Kouloumpis, T. Wilson, J. Moore, "Twitter sentiment analysis: the good, the bad and the OMG!", *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011

[49] D. Garbugli, "Metodologia per l'aggiornamento continuo di una rete semantica orientata alla sentiment analysis nei social media", *Master of Science degree thesis,* 2011

[50] L. C. Carminati, "Una metodologia per la valutazione del sentiment web basata sul concetto di reputazione", *Master of Science degree thesis,* 2010

[51] B. Liu, M. Hu, J. Cheng, "Opinion observer: analyzing and comparing opinions on the web", *Proceedings of the 14th international conference on World Wide Web (ACM)*, pages 342–351, 2005

[52] I. Anger, C. Kittl, "Measuring influence on twitter", *Proceedings of the 11th International Conference on Knowledge Management and*

Knowledge Technologies (ACM), page 31, 2011

[53] S. Asur, B. A. Huberman, G. Szabo, C. Wang, "Trends in social media: Persistence and decay", *5th International AAAI Conference on Weblogs and Social Media*, 2011

[54] Y. Ni, Z. Liu, "Heuristic search for optimizing diffusion of influence in a social network under the resource constraint", *Springer-Verlag*, 2010

[55] K. Larson, R. T. Watson, "The value of social media: toward measuring social media strategies", *Proceedings of the Thirty Second International Conference on Information Systems (ICIS)*, 2011

[56] Wikipedia entry for sentiment analysis
http://en.wikipedia.org/wiki/Sentiment_analysis

[57] Wikipedia entry for Klout
http://en.wikipedia.org/wiki/Klout

[58] Official Klout website
www.klout.com

[59] Wikipedia entry for Twitter
http://en.wikipedia.org/wiki/Twitter

[60] Official Twitter website
www.twitter.com

[61] L. Bruni, "A methodological framework to understand and leverage the impact of content on social media influence", *Doctoral dissertation*, 2013

[62] L. Bruni, C. Francalanci, P. Giacomazzi, F. Merlo, A. Poli, "The relationship among volumes, specificity, and influence of social media information", *Thirty Fourth International Conference on Information Systems,* 2013

[63] G. Anburaj, A. D. Rajkumar, "The art of public speaking", *Radix international journal of research in social science*, 2012

[64] D. O'Hair, R. Stewart, H. Rubenstein, "A Speaker's Guidebook: Text and Reference", *Macmillan Higher Education*, 2009

[65] Arthur Koch, "Speaking with a purpose", *Prentice-Hall*, 1988

[66] V. Kumar, "The Art Of Public Speaking", *Sterling Publishers Pvt. Ltd*, 2005

[67] L. Alfino, "*Know your audience*", *Writer*, Volume 116, Issue 5, page 38, 2003

[68] J. Vitak, P. Zube, A. Smock, C. T. Carr, N. Ellison, C. Lampe, "It's complicated: Facebook users' political participation in the 2008 election", *CyberPsychology, behavior, and social networking*, Volume 14, Issue 3, pages 107–114, 2011

[69] M. L. Knapp, J. A. Hall, "Nonverbal communication in human interaction", *Wadsworth Publishing Company*, 2009

[70] J. Horrigan, J. Boase, L. Rainie, B. Wellman, "The strength of internet ties", *Pew Internet & American Life Project report,* pages 33–37, 2006

[71] E. Fischer, A. R. Reuber, "Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior?", *Journal of Business Venturing*, 2011

[72] J. Busch, "Once is not enough - selling frequency", *Link & Learn*, 2010

[73] K. W. Wilson, A. Divakaran, "Broadcast video content segmentation by supervised learning. In Multimedia Content Analysis", *Springer*, pages 1–17, 2009

[74] J. Berger, K. Milkman, "Social transmission, emotion, and the virality of online content", *Wharton Research Paper*, 2010

[75] B. J. Jansen, M. Zhang, K. Sobel, A. Chowdury, "Twitter power: Tweets as electronic word of mouth", *Journal of the American society for information science and technology*, Volume 60, Issue 11, pages 2169–2188, 2009

[76] M. H. Ruge, J. Galtung, "The structure of foreign news. The presentation of the Congo, Cuba, and Cyprus crises in four norwegian newspapers", *Journal of Peace Research*, Volume 2, Issue 1, pages 64–91, 1965

[77] S. Polvara, "Studio empirico delle dinamiche di diffusione dell'informazione in Twitter", *Master of Science degree thesis,* 2012

[78] D. Barbagallo, L. Bruni, C. Francalanci, P. Giacomazzi, "An empirical study on the relationship between sentiment and influence in the tourism domain", *19th eTourism community conference (ENTER2012)*, pages 506–516, 2012

[**79**] Official Twitter4j website
www.twitter4j.org

[**80**] Klout4j repository
http://code.google.com/p/klout4j/

[81] L. Bruni, "Sviluppo di una metodologia per la qualità delle analisi di web sentiment basata su tecniche di data cleaning", *Master of Science degree thesis,* 2010

[82] S. Anholt, "Competitive identity: the new brand management for nations, cities and regions", *Palgrave MacMillan,* 2007

[**83**] A. Bruns, S. Stieglitz, "Towards more systematic Twitter analysis: metrics for tweeting activities", *International Journal of Social Research Methodology*, page **6**, note **7**, **2013**

[84] "The structure of the Java Virtual Machine – Runtime data areas – Heap", Oracle Java SE7 official documentation, cap. 2
http://docs.oracle.com/javase/specs/jvms/se7/html/jvms-2.html

[85] MySQL 5.5 Reference Manual (SELECT syntax), see
http://dev.mysql.com/doc/refman/5.0/en/select.html

[86] Stack Overflow discussion about limit keyword
http://stackoverflow.com/questions/4481388/why-does-mysql-higher-limit-offset-slow-the-query-down

[87] Twitter official support documentation – retweets
http://support.twitter.com/articles/77606-faqs-about-retweets-rt#

[88] Twitter official developers documentation – streaming APIs

limitations http://dev.twitter.com/docs/faq#6861

[89] Twitter official support documentation – tweet length http://support.twitter.com/articles/78124

[90] Twitter official support documentation – mentions http://support.twitter.com/articles/14023#

[91] Twitter official support documentation – hashtags and mentions troubleshooting http://support.twitter.com/articles/370610-my-hashtags-or-replies-aren-t-working

[92] Twitter official developers documentation – characters limit http://dev.twitter.com/docs/counting-characters

[93] L. Radice, S. Bruno, "Analisi della reputazione in ambienti Web 2.0: dallo studio alla realizzazione di un tool per il Sentiment Analysis", *Master of Science degree thesis*, 2009

[94] Wikipedia entry for Web 2.0 http://en.wikipedia.org/wiki/Web_2.0

[95] Wikipedia entry for Ajax http://en.wikipedia.org/wiki/Ajax_(programming)

[96] Wikipedia entry for Javascript http://en.wikipedia.org/wiki/JavaScript

[97] Wikipedia entry for Social Networks http://en.wikipedia.org/wiki/Social_network

[98] Wikipedia entry for Graph Theory

http://en.wikipedia.org/wiki/Graph_theory

[99] Wikipedia entry for Social Media
http://en.wikipedia.org/wiki/Social_media

[100] Wikipedia official website
http://en.wikipedia.org

[101] Facebook official website
www.facebook.com

[102] LinkedIn official website
www.linkedin.com

[103] Wikipedia entry for Wisdom of the Crowd
http://en.wikipedia.org/wiki/Wisdom_of_the_crowd

[104] Wikipedia entry for Microblogging
http://en.wikipedia.org/wiki/Microblogging

[105] Sean Golliher website, research article on reversing Klout score
http://www.seangolliher.com/2011/uncategorized/how-i-reversed-engineered-klout-score-to-an-r2-094/

[106] Official Time 100 charts website
http://time100.time.com/

[107] Wikipedia entry on Time 100
http://en.wikipedia.org/wiki/Time_100#Multiple_appearances

[108] Official Esquire website
http://www.esquire.com/features/most-influential-21st-century-1008

[109] FanPageList official website
http://fanpagelist.com/

[110] Klout official developers documentation
http://klout.com/s/developers/v2#scores

[111] A. Agresti, B. Finlay, "Statistical Methods for the Social Sciences", *Pearson*, 2008

[112] M. E. J. Newman, "Power laws, pareto distributions and zipf's law", *Contemporary physics*, Volume 46, Issue 5, pages 323–351, 2005

[113] R. B. Kline, "Principles and practice of structural equation modeling", *The Guilford Press*, 2010

[114] P. M. Bentler, D. G. Bonett, "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures", *Psychological bulletin*, Volume 88, Issue 3, pages 588 – 606, 1980

[115] K. G. Jöreskog, D. Sörbom, "Lisrel 8: Structural Equation Modeling with the Simplis Command Language", *Scientific Software*, 1993

[116] R. E. Schumacker, R. G. Lomax, "A Beginner's Guide to Structural Equation Modeling", *Lawrence Erlbaum Associates*, 2004

[117] K. A. Bollen, J. S. Long, "Testing Structural Equation Models", *Sage Focus Edition*, 1993

[118] L. T. Hu, P. M. Bentler, "Cutoff Criteria for Fit Indexes in

Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, Volume 6, Issue 1, pages 1 – 55, 1999

[119] L. R. Tucker, C. Lewis, "A Reliability Coefficient for Maximum Likelihood Factor Analysis," *Psychometrika*, Volume 38, Issue 1, pages 1 – 10, 1973

[120] P. Bentler, "Comparative Fit Indexes in Structural Models", *Psychological bulletin*, Volume 107, Issue 2, pages 238 – 246, 1990

[121] B. M. Byrne, "Structural Equation Modeling with Lisrel, Prelis, and Simplis: Basic Concepts, Applications, and Programming", *Psychology Press*, 1998

[122] J. H. Steiger, "Understanding the Limitations of Global Fit Assessment in Structural Equation Modeling," *Personality and Individual Differences*, Volume 42, Issue 5, pages 893 – 898, 2007

[123] D. A. Kenny, B. Kaniskan, D. B. McCoach, "The performance of RMSEA in models with small degrees of freedom", *University of Connecticut*, 2011 & 2013

[124] J. Pallant, "SPSS survival manual: A step by step guide to data analysis using SPSS", *Open University Press*, 2010

[125] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space", *Philosophical Magazine,* Volume 2, Issue 11, pages 559–572, 1901

[126] R. P. Bagozzi, Y. Yi, "On the evaluation of structural equation models", *Journal of the academy of marketing science*, Volume 16,

Issue 1, pages 74–94, 1988

[127] C. Fornell, D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error", *Journal of marketing research*, pages 39–50, 1981

[128] J. Cohen, "Statistical power analysis for the behavioral sciences", *Psychology Press*, 1988

[129] H. F. Kaiser, "An Index of Factorial Simplicity", *Psychometrika*, Volume 39, Issue 1, 1974

[130] S. Bolasco, "Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione", *Carocci*, 1999

[131] H. F. Kaiser, "The application of electronic computers to factor analysis", *Educational and Psychological Measurement*, Volume 20, 1960

[132] Official IBM Amos 20.0 user guide ftp://public.dhe.ibm.com/software/analytics/spss/documentation/amos/20.0/en/Manuals/IBM_SPSS_Amos_User_Guide.pdf

[133] R. P. Bagozzi, C. Fornell, "Theoretical concepts, measurements, and meaning - A second generation of multivariate analysis: measurement and evaluation", *Praeger*, pages 24–38, 1982

[134] B. Wheaton, B. Muthen, D. Summers Alwin, "Assessing reliability and stabillity in panel models", *Sociological Methodology*, 1977

[135] L. Hayduk, G. Cummings, K. Boadu, H. Pazderka-Robinson, S.

Boulianne, "Testing! Testing! One, two, three-testing the theory in structural equation models!", *Personality and Individual Differences*, Volume 42, Issue 5, pages 841–850, 2007

[136] M. W. Browne, R. Cudeck, "Alternative ways of assessing model fit", *K. A. Bollen & J. S. Long editors*, 1993

[137] R. Nuzzo, "Scientific method: Statistical errors", *Nature*, Volume 506, Issue 7487, page 150, 2014

[138] V. E. Johnson, "Revised standards for statistical evidence", *Proceedings of the National Academy of Sciences of the United States of America*, 2013

[139] J. C. Beck, T. H. Davenport, "The attention economy: Understanding the new currency of business", *Harvard Business Press*, 2001

[140] Article on the blog of a Chinese PhD student, which summarizes chi-square acceptance conditions and fit indexes ranges in AMOS http://zencaroline.blogspot.it/2007/04/global-model-fit.html