TÉCNICO LISBOA

Influence and Sentiment Homophily on Twitter

Hugo Manuel Antunes Lopes

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Helena Sofia Andrade Nunes Pereira Pinto
Prof. Alexandre Paulo Lourenço Francisco

Examination Committee

Chairperson: Prof. João António Madeiras Pereira
Supervisor: Prof. Helena Sofia Andrade Nunes Pereira Pinto
Members of the Committee: Prof. Bruno Emanuel da Graça Martins

November 2015

Acknowledgments

Podia desenhar um pontinho por cada pessoa que me acompanhou neste trajecto, seria como olhar à noite o céu limpo com as estrelas e o seu diferente brilho, consoante estejam mais longe ou mais perto. Podia traçar uma linha entre cada par de pontos conhecidos, desenhava constelações de amigos e iria encontrar pessoas que me guiaram no caminho mesmo sem eu saber.

Este trabalho não existiria sem a preciosa ajuda da Professora Sofia e do Professor Alexandre e de quem dispensou um pouco do seu tempo para contribuir para avaliação dos resultados obtidos, bem como de quem me deu conselhos, sugeriu ideias ou me distraiu durante este ano.

Aos meus pais que tudo fizeram para que eu tivesse a oportunidade de chegar até aqui, ao meu irmão, à minha namorada, à minha família, aos meus amigos de sempre, aos meus amigos de faculdade, aos meus amigos de agora, aos meus amigos que não vejo a todas as horas e a todos os que me trouxeram até aqui, um muito sincero e enorme obrigado.

Abstract

The Web has revolutionized democratization of knowledge and it is now democratizing our social relationships through several social media websites, like Twitter. These On-line Social Networks have millions of users, widely connected, who communicate and interact in an unparalleled dynamic level. Twitter not only connects people, it is also a window for their interactions. We can collect data about social networks and their dynamics from Twitter, represent them and reasoning about them. The inherent sentiment of these interactions and phenomenons such as influence are observable, but not easily inferable, and, with this work, we aim to understand if Influence and Sentiment are correlated. We present an empirical study that combines existing Graph Clustering and Sentiment Analysis techniques for reasoning about Sentiment dynamics at cluster level and analyzing the role of Social Influence on Sentiment contagion, based on a large dataset extracted from Twitter during the 2014 FIFA World Cup. Exploiting WebGraph and LAW frameworks to extract clusters, and SentiStrength to analyze sentiment, we propose a strategy for finding moments of Sentiment Homophily in social circles. We found that clusters tend to be neutral for long ranges of time, but denote volatile bursts of sentiment polarity locally over time. In those moments of polarized sentiment homogeneity, there is evidence of an increased, but not strong, chance of one sharing the same overall sentiment that prevails on the community to which he belongs.

Keywords: Social Networks, Twitter, Social Circles, Influence, Sentiment Homophily, 2014 FIFA World Cup.

Resumo

A Internet revolucionou a democratização do conhecimento e assistimos agora à democratização das nossas próprias relações sociais através diferentes serviços de redes sociais na rede, como o Twitter. Estas redes sociais têm milhões de utilizadores ligados por todo o mundo, que interagem num ritmo sem precedentes. O Twitter não só liga as pessoas entre si como é também uma janela aberta para as suas interacções, sendo possível recolher, representar e analisar informação sobre estas redes sociais. O sentimento intrínseco a estas interacções e fenómenos como a influência são observáveis nas relações entre as pessoas e, por vezes, estão inerentes à forma como se alteram ou evoluem, mas não são facilmente inferíveis. Com este trabalho pretendemos perceber se a Influência e o Sentimento estão correlacionados. Aqui apresentamos um estudo empírico que combina técnicas para encontrar comunidades e de análise de sentimento para analisar a dinâmica geral do sentimento num círculo social, usando um conjunto de dados extraído do Twitter durante o Mundial de Futebol de 2014. Tirando partido das ferramentas WebGraph e LAW para encontrar círculos sociais e analisando o sentimento através do SentiStrength, nós propomos uma estratégia para encontrar momentos de homofilia de sentimento nessas comunidades. Com este trabalho descobrimos que as comunidades tendem a apresentar longos períodos de neutralidade intercalados com momentos de polarização de sentimento. Quando nesses momentos existe homogeneidade de sentimento, há uma maior probabilidade, embora não muito forte, de alguém pertencente a esse círculo social partilhar um sentimento equivalente àquele que prevalece na comunidade.

Palavras-Chave: Redes Sociais, Twitter, Círculos Sociais, Influência, Homofilia de Sentimento, Campeonato do Mundo de Futebol no Brasil 2014.

Contents

List of Tables **xi**

List of Figures **xv**

1 Introduction **3**
1.1 Motivation . . . 3
1.2 Hypotheses . . . 4
1.3 Objectives . . . 6
1.4 Results Summary . . . 7
1.5 Organization . . . 7

2 Related Work **9**
2.1 Social Networks in Theory: A generic overview . . . 9
2.1.1 Graphs as a representation of Networks . . . 9
2.1.2 Centrality Measures . . . 13
2.1.3 Tie Strength and Network's Dynamic . . . 14
2.1.4 The Leading Role of Weak Ties . . . 15
2.1.5 Power and Place in the Network . . . 17
2.1.6 Popularity Models . . . 18
2.1.7 Relationship Polarity and Network's Shape . . . 19
2.1.8 Social Similarity and Context Surrounding Influence . . . 21
2.1.9 Influence and Information Cascades . . . 22
2.1.10 Influence and Cascading Behavior . . . 25
2.1.11 Information Diffusion and Epidemics . . . 27
2.2 Twitter: A Wide Social Environment . . . 28
2.2.1 Tie Strength on Twitter . . . 30
2.2.2 Network Structure and Finding Communities . . . 30
2.2.3 Event Detection . . . 31
2.2.4 Event Prediction . . . 32
2.2.5 Information Flow . . . 33
2.2.6 Influence and Homophily . . . 34
2.2.7 Sentiment Analysis: Positivity, Negativity, Neutrality . . . 34
2.2.8 Spam Filtering . . . 36
2.2.9 Geo-location . . . 37
2.2.10 2010 FIFA World Cup on Twitter . . . 38
2.2.11 Twitter as a mirror for other Social Environments . . . 38
2.3 Combining Community Detection, Sentiment Analysis, Influence and Homophily . . . 39
2.4 Summary . . . 40

3 Data Overview **43**
3.1 Twitter Developer APIs . . . 43
3.2 Extracted Dataset . . . 44

4 Approach **49**
4.1 User Clustering . . . 49
4.2 Tweet Clustering . . . 51
4.3 Sentiment Analysis . . . 52
4.4 Influence and Sentiment Homophily Analysis over Time . . . 54
4.4.1 Sentiment Homophily in Narrow Time Clusters . . . 54
4.4.2 Polarity Changes in Wide Time Clusters . . . 55
4.4.3 Local Sentiment Homophily in Wide Time Clusters . . . 56

5 Results **61**
5.1 User Clustering . . . 61
5.2 Tweet Clustering . . . 68
5.3 Influence and Sentiment Homophily Analysis over Time . . . 73
5.3.1 Sentiment Homophily in Narrow Time Clusters . . . 73
5.3.2 Polarity Changes in Wide Time Clusters . . . 74
5.3.3 Local Sentiment Homophily in Wide Time Clusters . . . 84

6 Evaluation **87**
6.1 Modularity Measure for User Clustering . . . 87
6.2 Manual validation for Local Polarity Homophily . . . 89
6.3 Manual validation of non-ambiguous classifications . . . 91
6.4 Krippendorff's alpha reliability about Human-coders Agreement . . . 92
6.5 K-fold Cross Validation . . . 93

7 Conclusion **95**

8 Future Work **97**

Bibliography **99**

A Twitter Data Keywords **107**

B Structure of JSON encoded Tweets **109**

C Sentiment Polarity Classification for Human-coders **113**

List of Tables

1.1 Summary of different social media environments, according to their most recent official statistics. . . . 4

2.1 Payoff matrix of *w* and *v* choosing behavior A or B . . . 25
2.2 SentiStrength evaluation results for Twitter data [89]. Metric used: accuracy regarding the golden standard created by 3 human coders. Comparison between Unsupervised and Supervised SentiStrength and the best result of different machine learning techniques used. . . . 36
2.3 Relevant contributions suitable to the scope of this work, with comparison between some different techniques and approaches. In bold are represented the ideas and methodologies that we chose to follow in our research. . . . 42

3.1 Tweet type distribution. . . . 45
3.2 Tweet type distribution in the knock-out stage subset. . . . 45

5.1 Summary of Global User Clustering characteristics. . . . 61
5.2 Summary of Daily-based User Clustering characteristics per day for **retweets** graph, with information about the games schedule. . . . 62
5.3 Summary of Daily-based User Clustering characteristics per day for **replies** graph, with information about the games schedule. . . . 63
5.4 Summary of Round-based User Clustering characteristics for **retweets**. . . . 63
5.5 Summary of Round-based User Clustering characteristics for **replies**. . . . 63
5.6 Comparison between the number of clusters of users and the number of clusters of tweets, obtained with daily-based clustering. The differences between retweets and replies in each language are also compared. . . . 68
5.7 Comparison between the number of clusters of users and the number of clusters of tweets, obtained with round-based clustering. The differences between retweets and replies in each language are also compared. . . . 68
5.8 Number of completely neutral clusters and the number of clusters with polarity spikes for round-based clusters. . . . 76
5.9 Comparison between the number of clusters with size equal or greater than 10 and 100 and the number of clusters that have ambiguous sentiment classifications in periods of sentiment homophily, for each different strategy used. . . . 85

6.1 Modularity results for each set of clusters. . . . 88
6.2 Manual evaluation results regarding the approach implemented in the Algorithm 4, and comparison with a random approach. . . . 91
6.3 Manual evaluation results of polarized sentiment classifications obtained with SentiStrength. 92

6.4 Error rate E average of K-Fold Cross Validation, for $k = 10$, over sets of tweets in periods of prevalence of a certain sentiment polarity. 94

List of Figures

1.1 Example of a tweet posted by National Aeronautics and Space Administration – United States of America (NASA), using the hashtags #WorldCup and #Brazil, which was retweeted by 806 users. 5

2.1 Representation of my personal Twitter Network. Nodes in red represent User Accounts, and whenever there are mutual Follow relations these are represented by the edges in grey. 10
2.2 Representation of my personal Twitter Network of Followers. Nodes in red represent User Accounts, and directed Follow relations are represented by the edges in grey. 11
2.3 Example of a Breadth-first Search starting at my personal Twitter account. Layer 0 is the root at a distance of 0. Layer 1 represents root Followers at a distance of 1. Layer 2 represents Followers of three arbitrary root Followers at a distance of 2. 12
2.4 Node 1 has three neighbors: 18, 29, 30. From the three possible pairs of neighbors, only the pair (18, 30) is also connected, which gives $CC_G(1) = \frac{1}{3}$. 15
2.5 When two individuals have a common friend, they are probably aware of each other and there is an increased chance of being friends in the future. 16
2.6 Every edge in this figure is a bridge. Whichever edge that could be removed, it separates its endpoints into two independent components. 17
2.7 The local bridge highlighted in green decreases the distance between the two densely connected components to which it is attached (1 and 3). Without this edge, the path between these two components would cross another component (2). The node highlighted in blue is a structural hole because it is connected to two densely connected components (3 and 4), through two local bridges. 18
2.8 Illustration of Structural Balance [82], (I) and (II) are balanced, while (III) and (IV) are not. 19
2.9 A balanced graph with both positive and negative relations can be divided in two groups, (X) and (Y), in such away that all the edges inside each group are positive and all the edges between the two groups are negative, according to the Balance Theorem. 20
2.10 Illustration of Social Status Theory [82]. (I) and (II) satisfy Status Theory, while (III) and (IV) do not. 21
2.11 Example of Triadic Closure in Twitter [69]. User *i* follows *j*, who follows *k*. When *k* posts a tweet and *j* retweets it, *i* gets exposed to the tweet and there is an increased likelihood of *i* to follow *k*. 31

3.1 Scheme of the HTTP connection to Twitter Public Streaming API [100] 44
3.2 Presence of tweets with Uniform Resource Locator – web address (URL) in the dataset. . 45
3.3 Top 10 languages in the dataset: English, Spanish, Portuguese, Indonesian, French, Japanese, Italian, German, Turkish, Dutch. 46
3.4 Top 10 languages in knock-out stage subset: English, Spanish, Portuguese, Indonesian, French, Japanese, Italian, German, Turkish, Dutch. 47

4.1 High-level view of the designed workflow. 49
4.2 User clustering process. 51
4.3 Tweet clustering process. 53

5.1 Distribution of clusters, obtained from Global User Clustering, regarding their size. On top, the comparison between the totality of **retweet-based** clusters distribution and the distribution of those who belong to the giant component. On bottom, the same comparison for replies. The exponent value of each power-law is, respectively, 1.87512, 1.86919, 2.02154, and 2.02124. 65
5.2 Distribution of clusters, obtained from Daily-based Clustering for July 9th, regarding their size. Comparison between the totality of **retweet-based** clusters distribution and the distribution of those who belong to the giant component. The exponent value of each power-law is, respectively, 1.90412 and 1.90028. 66
5.3 Distribution of clusters, obtained from Daily-based Clustering for July 10th, regarding their size. Comparison between the totality of **reply-based** clusters distribution and the distribution of those who belong to the giant component. The exponent value of each power-law is, respectively, 2.21169 and 2.18112. 66
5.4 Distribution of clusters, obtained from Round-based User Clustering for the Semi-finals, regarding their size. On top, the comparison between the totality of retweet-based clusters distribution and the distribution of those who belong to the giant component. On bottom, the same comparison for replies. The exponent value of each power-law is, respectively, 1.89057, 1.88808, 2.15327, and 2.14271. 67
5.5 Distribution of clusters, obtained from Daily-based Tweet Clustering for **retweets** on July 13th, regarding their number of tweets. 69
5.6 Distribution of clusters, obtained from Daily-based Tweet Clustering for **replies** on July 13th, regarding their number of tweets. 70
5.7 Distribution of clusters, obtained from Round-based Tweet Clustering for **retweets** on the Final stage, regarding their number of tweets. 71
5.8 Distribution of clusters, obtained from Round-based Tweet Clustering for **replies** on the Final stage, regarding their number of tweets. 72
5.9 Distribution of absolute sentiment values and sentiment pair values in Spanish-speaking cluster “334972” on July 13th. Cluster size: 23 tweets. 74
5.10 Distribution of absolute sentiment values and sentiment pair values in Portuguese-speaking cluster “797328” on July 13th. Cluster size: 222 tweets. 75
5.11 Comparison between the overall distribution results for the four languages on July 8th, considering retweet-relation clusters. 78
5.12 Time-line of tweets’ frequency of absolute sentiment for each accumulation of 3 hours. Cluster “413547” from the Spanish-speaking set of **reply-based** clusters over the Quarter-finals stage. 79
5.13 Time-line of tweets’ frequency of absolute sentiment for each accumulation of 3 hours. Cluster “5171567” from the English-speaking set of **retweet-based** clusters over the Round of 16. 79
5.14 Time-line of tweets’ frequency of absolute sentiment for each accumulation of 3 hours. Cluster “553712” from the Portuguese-speaking set of **retweet-based** clusters over the Semi-finals stage. 79

5.15 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "301837" from the English-speaking set of **reply-based** clusters over the Round of 16. 80
5.16 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "2049176" from the Spanish-speaking set of **retweet-based** clusters over the Final stage. 80
5.17 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "2479319" from the Portuguese-speaking set of **retweet-based** clusters over the Final stage. 81
5.18 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "1911770" from the German-speaking set of **retweet-based** clusters over the Semi-finals stage. 82
5.19 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "1000883" from the English-speaking set of **reply-based** clusters over the Semi-finals stage. 83
5.20 Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "177613" from the Portuguese-speaking set of **reply-based** clusters over the Final stage. 83
5.21 Example of an ambiguity, with the sentiment value of (1, -1), surrounded by a negative context. 85

C.1 Classification environment for human-coders evaluate the sentiment polarity of the displayed tweet as positive, neutral, or negative. 114
C.2 Classification environment for human-coders evaluate the sentiment polarity of the displayed tweet as positive, neutral, or negative. 114

Chapter 1

Introduction

1.1 Motivation

People are connected to each other since the earliest times of the first human tribes. So long as primitive men started to interact and communicate among them, humans started creating ties with each other. Sets of these ties are called Social Networks. In the 1950s the field of Social Networks was born having the goal of conceptualizing these connections and building tools to represent the networks inherent to these human interactions [104]. The term Social Networks invaded our lives with the rise of several web-based social network services. The way we relate to each other and the means we use to communicate are several, diversified, and they are continuously evolving. The Web created a new paradigm of communication, and Social Media plays an emergent role in how people interact today. Social Networks are on everyone's lips and in everyone's pocket, their increasing popularity encourages more and more people to participate in various online activities, which produce data at an unprecedented rate [80]. But the Web not only acts as a new channel for people's interactions, mirroring their dynamic network, it is also a huge source of data that can be used to represent and study these Social Networks. These representations of our connections with other people allow us to reason about our relationships [104].

Facebook [1], Twitter [2], Instagram [3], Pinterest [4], LinkedIn [5], Google+ [6], Tumblr [7] are examples of the most popular social media websites today. As it is described in Table 1.1, they are all used by millions of people with Facebook, Twitter and Instagram in the lead of daily-usage, being used all over the world and wide language coverage. However, each one of them has its own characteristics and different purposes that have lead the different social environments to give special emphasis to particular forms of interaction. For instance, Facebook is designed to extend friendship relations to the web, LinkedIn is intended to reinforce professional networks, Twitter aims to facilitate information sharing in real-time, Instagram encourages people to share their moments in pictures or videos, and this lead to different type activities. Despite these differences, they all promote and facilitate the connection between individuals creating social networks.

Twitter and Tumblr are known as microblogging platforms, due to their short-content interactions (based on short-messages broadcast) [37, 17]. Our work is based on Twitter, which is widely used by its 316 million monthly active users, who generate about 500 million messages per day [95]. This message

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://www.instagram.com/
[4] https://www.pinterest.com/
[5] https://www.linkedin.com/
[6] https://plus.google.com/
[7] https://www.tumblr.com/

Social Network	**Facebook**	**Twitter**	**Instagram**	**Pinterest**	**LinkedIn**	**Google+**	**Tumblr**
Founded	2004	2006	2010	2010	2003	2011	2007
About	Stay connected with friends and family to discover what is going on in the world, and share and express what matters to you. [27]	Twitter is your window to the world. Get real-time updates about what matters to you. Create and share ideas and information instantly, without barriers. [95]	Capture and share the world's moments. Take a picture or video, then customize it with filters and creative tools. [46]	Pinterest is a visual bookmarking tool that helps you discover and save creative ideas. [73]	The world's largest professional network. You get access to people, jobs, news, updates, and insights that help you be great at what you do. [63]	A social network created for businesses. Google+ makes it faster and easier to share and collaborate with your customers and team members. [35]	Share anything effortlessly. Post text, photos, quotes, links, music and videos. You can customize everything. [93]
Users	1.49 billion (monthly active) 968 million (daily active) [27]	316 million (monthly active) [95]	300 million (monthly active) [47]	78 million (monthly active) [75]	380 million [61]	300 million [34]	249.8 million [94]
Activity	350 million photos uploaded per day 4+ billion video views per day 16 million events created per month [28]	500 million tweets sent per day [95]	30+ billion photos shared 2.5 billion daily likes 70 million photos shared per day [47]	50+ billion pins 3+ million pins sent per day [72]	130 000 posts per week [62]	1.5 billion photos shared per week [34]	74.1 million daily posts [94]
Accounts outside US	83% (and Canada) [27]	77% [95]	70% [47] +	-	62% [61]	-	58% [94]
Languages	70+ [28]	35+ [95]	25+ [47]	30+ [74]	24 [61]	60 [36]	16 [94]
Employees	10,955 [27]	4,100 [95]	-	500+ [74]	8 700 [61]	-	322 [94]

Table 1.1: Summary of different social media environments, according to their most recent official statistics.

exchange is similar to a Short Message Service (SMS) broadcast service that works on the Web [20]. People post and read 140-character messages, called *tweets* like Figure 1.1. Users can *follow* and be *followed* by other users. When some user *tweets* (sends a message), his/her *followers* are able to see it and *retweet* it, i.e., share that same tweet with their own followers, keeping the reference to the original author. Users can *mention* and *reply* to other users in their tweets and they can also attach photos, videos and links, as well as their geo-location. Another major feature is the use of *hashtags* that label the tweet with a set of topics, which somehow relate it with other tweets sharing the same hashtags, creating global conversations [64].

With this massive amount of data available in Twitter, not only Sociology and Psychology found it an invaluable source for study real interactions and dynamics in social networks, but also Politics, Marketing, Commerce, Civil Protection, Health Organizations found a great interest and potential in this stream of information [19, 53, 76, 9]. The applications vary [37] and we found in the literature different Social related Analysis, such as Network Dynamics [57, 45], Community Detection [66, 90], Event Detection and Prediction [76, 19, 18, 92], Information Flow [69, 10], Influence and Homophily Analysis [9, 16, 111, 81], Sentiment Analysis [41, 6]. However, dealing with this kind of data and extracting its intended meaning is not an easy task because these datasets are huge, linked, incomplete, highly dynamic, noisy and contain false information [80].

Community Detection techniques find social circles inside social networks, while Sentiment Analysis solutions infer the inherent sentiment of the texts that are exchanged in social interactions. Social Sciences observed that people tend to be similar to their peers, calling that *Homophily*, and they tend to adopt similar behaviors or characteristics of people to whom they are connected by *Influence*. The extent of characteristics that are commonly propagated through consequence of influence into a state of homophily is unknown and, it is plausible to think that social networks' topology may have a role in sentiment contagion.

1.2 Hypotheses

There are studies that analyze the interdependencies and possible correlations among different kinds of properties [12, 48, 90], however we found that there is not an extensive study about sentiment homo-

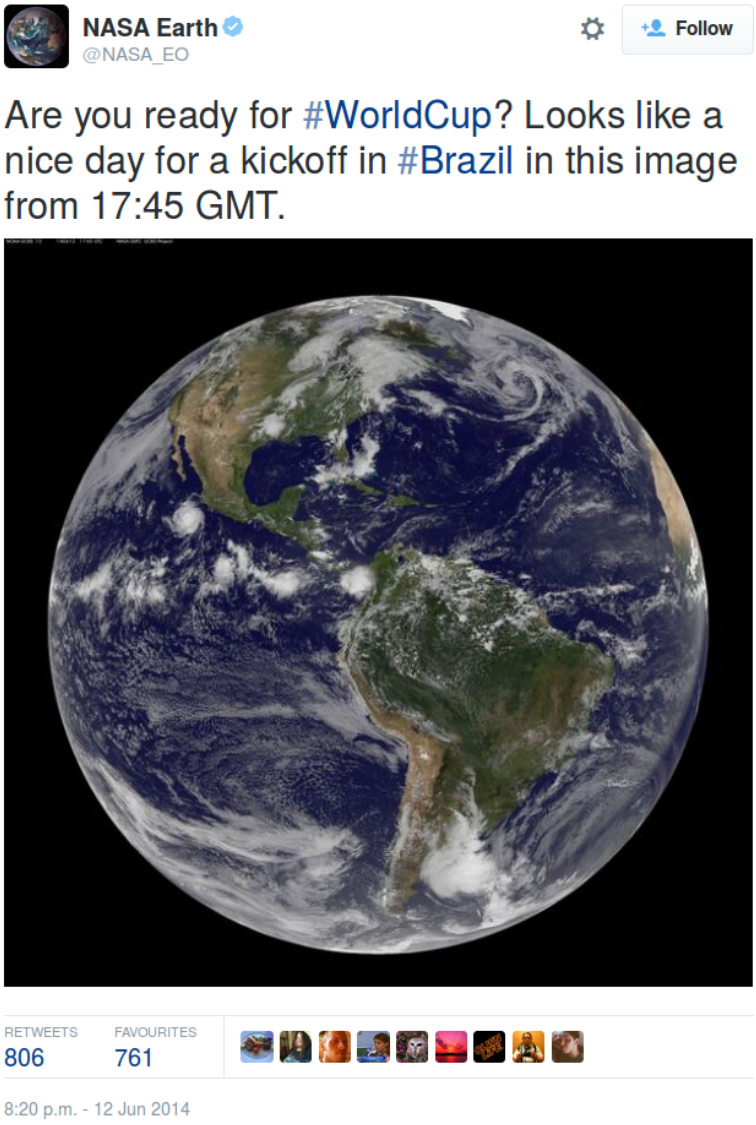


Figure 1.1: Example of a tweet posted by NASA, using the hashtags #WorldCup and #Brazil, which was retweeted by 806 users.

geneity in clusters and whether this sentiment can be propagated by influence into a state of sentiment homophily inside those clusters. Understanding how sentiment behaves at a cluster level can be useful for mining the overall mood of communities, and it may also be useful for improving sentiment classification techniques using enriched information about surrounding emotions. The hypotheses that motivate our work are:

- *H1:* The sentiment expressiveness inside clusters is highly dynamic over time.
- *H2:* Clusters show moments of sentiment prevalence.
- *H3:* During moments of sentiment homogeneity in a cluster, there is an increased chance that a user is influenced by the surrounding emotion and shows a similar sentiment to the one prevailing at that moment.

With these hypotheses we want to understand if it is viable to combine individual Sentiment Analysis with Social Theories of *Influence* and *Homophily* to make use of aggregated knowledge about social

communities, in order to classify a network's predominant sentiment, and also validate if the individuals' sentiment is inferable from the social circle to which he belongs.

1.3 Objectives

The period of time during which homophily is observable depends on the feature under analysis, but also depends on the time considered to find homophily and the threshold of homogeinuity defined to ascertain that individuals in the community tend to be similar regarding the considered feature. If we consider the feature to be the football club that individuals support in a social circle, it is possible that this characteristic remains static over time, with some residual changes. However, if we look for homophily regarding the brand of the mobile phone each individual owns it is quite possible that people will change it more often. In both cases it must be defined the minimum prevailing rate for an observation to be assumed that homophily exists. When this phenomenon is caused by influence, there is an increased chance of other individuals in the community to adopt the prevailing feature in the future [24]. On the other hand, Sentiment Analysis techniques still present a significant error rate, depending on the context and being especially fallible in sarcastic environments [89]. Having evidence about sentiment homophily between direct friends [85] we decided to analyze the behavior of sentiment at a community level, exploiting existing clustering techniques and searching for influence and homophily patterns. We aim to better understand the extent of influence and homophily and understand if they are related with sentiment. improve the sentiment analysis itself using contextual and aggregated information from the cluster.

To analyze sentiment dynamics in social circles, find moments of sentiment homophily, and understand the relation between sentiment homophily and influence, we had to combine sequentially Clustering and Sentiment Analysis techniques with Influence and Homophily Analysis. This process was divided into two major stages:

- Clustering and Sentiment Analysis;
- Influence and Sentiment Homophily.

With Clustering and Sentiment Analysis we combined existing techniques to find social circles and classify the inherent sentiment of their individuals' interactions (tweets), aiming to:

- Find clusters of users;
- Aggregate all tweets according to their users' clusters;
- Classify the sentiment of all tweets in each cluster.

Clustering and Sentiment Analysis was performed considering different periods of time of data to obtain different configurations of clusters with respective sentiment information. This data was used to analyze Influence and Sentiment Homophily, intending to:

- Observe sentiment dynamics of clusters;
- Systematically find moments of sentiment homophily;
- Understand whether the overall sentiment of a cluster during a period of sentiment homophily can be extrapolated for its individuals.

1.4 Results Summary

This work is based on a dataset of more than 339 million tweets about the 2014 International Federation of Association Football (FIFA) World Cup in Brazil, the biggest sporting event on Twitter in 2014 [96], where we found several different dynamic communities supporting different countries, from which we have selected only communities talking in English, Spanish, Portuguese and German. Using existing clustering and sentiment classification techniques, we propose to measure the overall sentiment of clusters based on the frequency of tweets for each possible sentiment value, regarding their sentiment classifications. We found that the neutral value is the most frequent classification during the clusters' time-life, however different sentiment values appear, usually in spikes and with different polarities over time, confirming the highly dynamic nature of clusters' sentiment (*H1*). We also observed moments of sentiment homophily (*H2*), for instance in chains of retweets or topic-related discussions and we describe a systematic strategy for finding those moments. Finally, we used dubious sentiment classifications for testing the role of influence in the origin of those moments of sentiment homophily by comparing the extrapolation of the clusters' overall sentiment with human-coders' evaluations. With this strategy we found a tendency for ambiguous classifications being correctly relabeled with the prevalent sentiment of its cluster (*H3*).

1.5 Organization

This report is structured as follows. In Chapter 2 we present the Related Work, introducing the standard theories and notions about Social Networks, with special attention to the subjects of Community Detection, Sentiment Analysis, Influence and Homophily, and how to extract them from Twitter. Then we describe the dataset in the basis of our work in Chapter 3. It is followed by Chapter 4 where we propose our solution for extracting information from the dataset to achieve the proposed objectives. We present the results and respective evaluation in Chapter 5 and 6. It ends with a reflection of the major findings and conclusions of this research in Chapter 7 and we discuss some different decisions and approaches that would be interesting to follow in the future in Chapter 8.

Chapter 2

Related Work

Today, with the advent of the Internet, Social Networks are a widely spread concept used in web-based social network services, which connect millions of users every day [24]. However, web-based services and the concept Social Network are often mixed together, because now the Web plays a central role in the way people interact and perceive their own social network [24]. In fact, the study of Social Networks began in the 1950s with strong connections to Sociology, Social Psychology and Anthropology [104]. The interest in this area has spread to many other areas such as Economics and Politics, and has its basis in Mathematics, Graph theory and Game theory.

In this chapter, we start by presenting the concepts and theories that are the basis of Social Networks in general, and then we introduce the most recent research work in this area, focusing on Twitter experiments, analyzing known problems, how they were solved and current open problems.

2.1 Social Networks in Theory: A generic overview

2.1.1 Graphs as a representation of Networks

Social Networks are composed by individuals and their relations. However, even though some of our relations are quite evident, they are not easily seen as whole. This leads to the first problem in the area: how can a social network be represented?

Graph Theory arises as a possible answer. The idea is to use graphs to simply and quickly describe a network. A *Graph* is the mathematical model of a network structure, represented by *Nodes* and *Edges* (another useful mathematical representation for networks is a matrix). *Edges* represent relations between *Nodes*. Nodes can be called vertices or endpoints, and edges are also commonly designated as links, ties, connections. Two *Nodes* are neighbors if they are connected by one *Edge*. *Graphs* can be directed, when there is an order in nodes' relations, or undirected, when the edge's node order is meaningless. Figure 2.1 and Figure 2.2 are examples of undirected and directed graphs that represent social networks, respectively.

Definition 1 (Graph) *A graph G is a pair of two sets $G(N, E)$ where N is a set of m nodes $N = \{n_1, n_2, ...n_m\}$ and E is a set of k edges $E = \{e_1, e_2, ..., e_k\}$. An edge e is a pair of nodes $e = (n_1, n_2)$ in which the order only matters when the graph is directed.*

A *path* in a graph is a sequence of nodes with the property that each consecutive pair is connected by an edge. There is a *cycle* when a path has at least three edges, and the first and the last nodes are the same, although all the other nodes are distinct. Cycles allow redundancy and alternative routes.

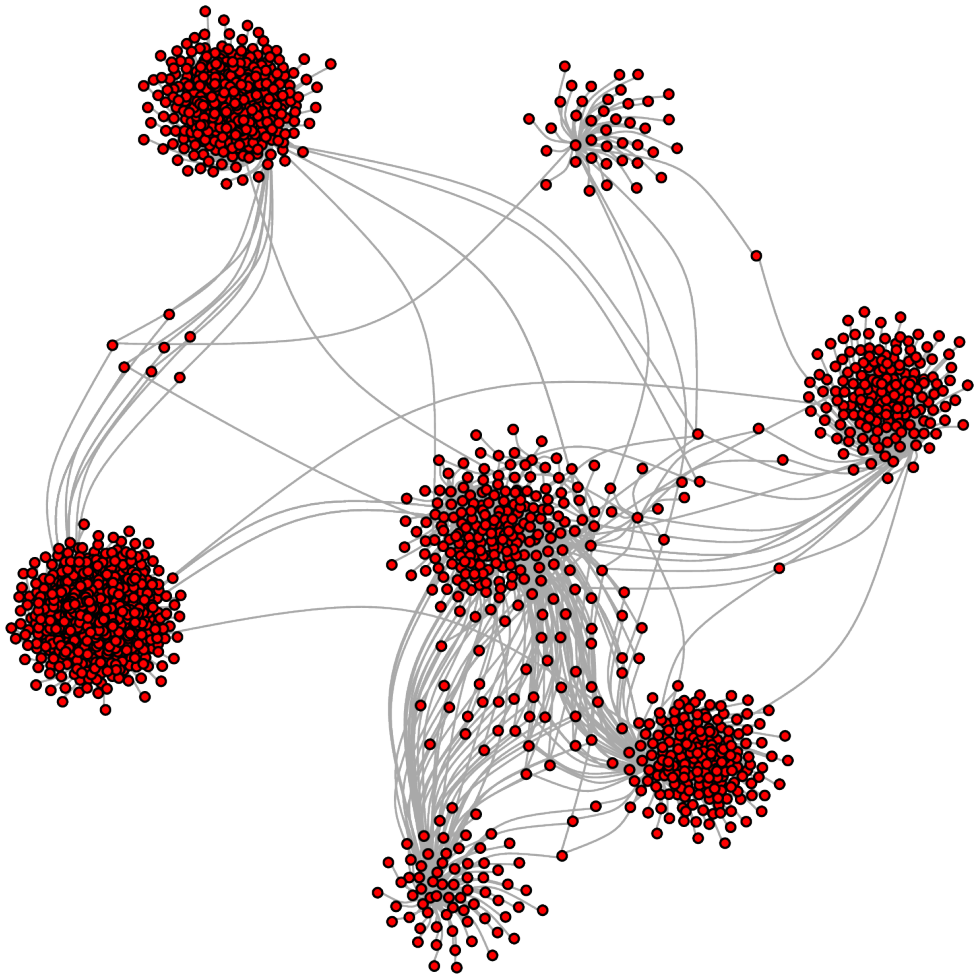


Figure 2.1: Representation of my personal Twitter Network. Nodes in red represent User Accounts, and whenever there are mutual Follow relations these are represented by the edges in grey.

Definition 2 (Path) *Path* $p = [n_x, n_y, ..., n_z]$ *with* k *nodes, where* $\forall x \in 0..k-1, (p[x], p[x+1]) \in E, (p[x-1], p[x]) \in E$. *If* $p[0] = n_x = p[k-1]$ *then the path is a cycle.*

The notion of *distance* between two nodes in a graph is defined as the length of the shortest path between them. The *length* is the number of edges in the sequence that comprises it.

Definition 3 (Length of Path) *Length* $l(p)$ *of path* p *with* k *nodes:*

$$l(p) = k \tag{2.1}$$

Definition 4 (Distance between two Nodes) *Distance between node* n_1 *and* n_2 *in a graph* G*, assuming* $L_{n_1 n_2}$ *as the set of all paths' length between* n_1 *and* n_2 *in* G*:*

$$d_G(n_1, n_2) = min(L_{n_1 n_2}) \tag{2.2}$$

However, measuring the distance can be difficult in some types of graphs, such as cyclic and large graphs. Therefore, it can require systematic approaches like a *Breadth-first Search*. This technique starts by transversing the graph from an initial node down to all its neighbors, such that the search is done by layers. To illustrate this, we can say that "me" is the starting node and it corresponds to the *layer 0*, all my friends are at the distance of 1 from me and they correspond to the *layer 1*, all the friends of

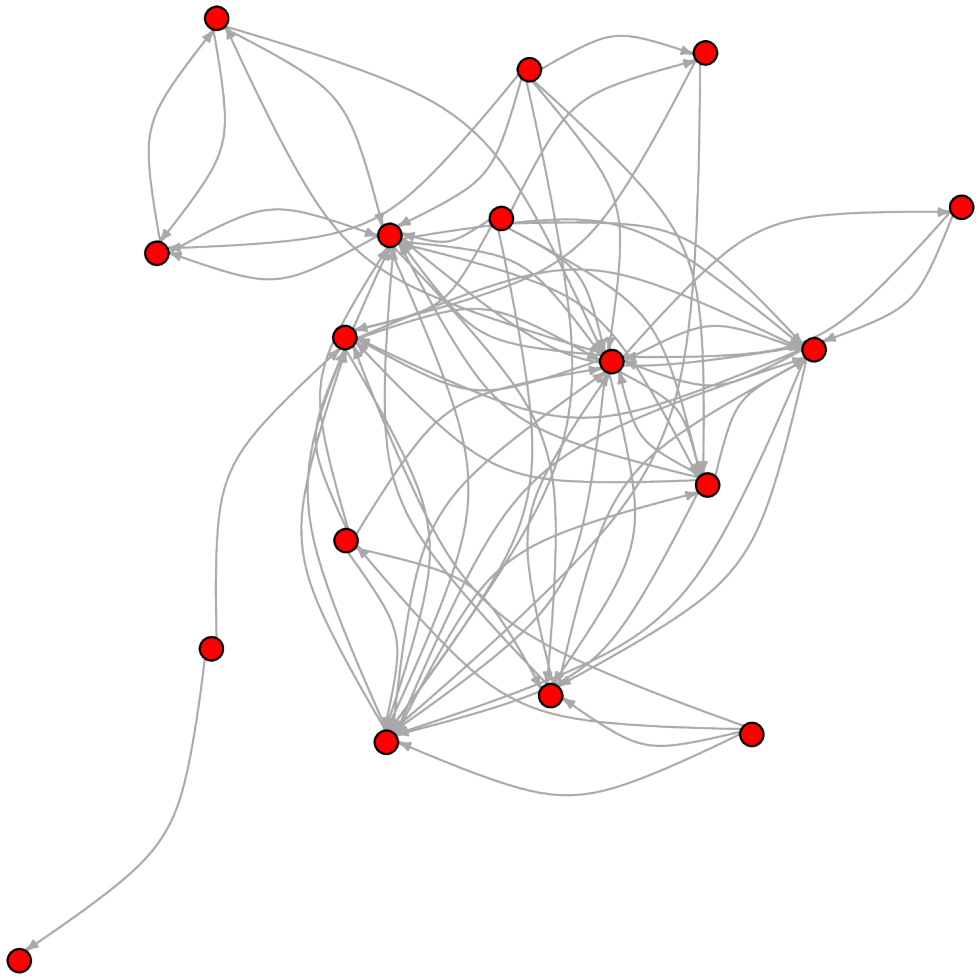


Figure 2.2: Representation of my personal Twitter Network of Followers. Nodes in red represent User Accounts, and directed Follow relations are represented by the edges in grey.

my friends are at the distance of 2 and they correspond to the *layer 2*, and so on, as it is represented in Figure 2.3.

A graph is *connected* if, for every pair of nodes, there is a path between them.

Definition 5 (Connected Component) *A* connected component *of a graph is a subset of nodes such that:* (i) every node in the subset has a path to every other. (ii) The subset is not part of some larger connected component.

Definition 6 (Single-node Component) *A node that does not belong to any pair of edges represents a* single-node component*.*

In order to build these graphs it is important to clearly define what a *node* and an *edge* represent. In a friendship network nodes may represent people and edges their friendship relations.

Considering the distance in the world friendship network, some experiments have shown that, not only there are paths of friends connecting you to a large fraction of the world's population, but also that these paths are surprisingly short - this is known as the Small-World Phenomenon. Furthermore, an experiment conducted by Stanley Milgram [91], using forwarding letters starting in randomly chosen people with some specific targets obtained a median length of six in the paths between the starters and their targets. More recently, the Leskovec and Horvitz experiment [56] with Microsoft Instant Messaging interactions, in a one month period, has shown an average distance of 6.6.

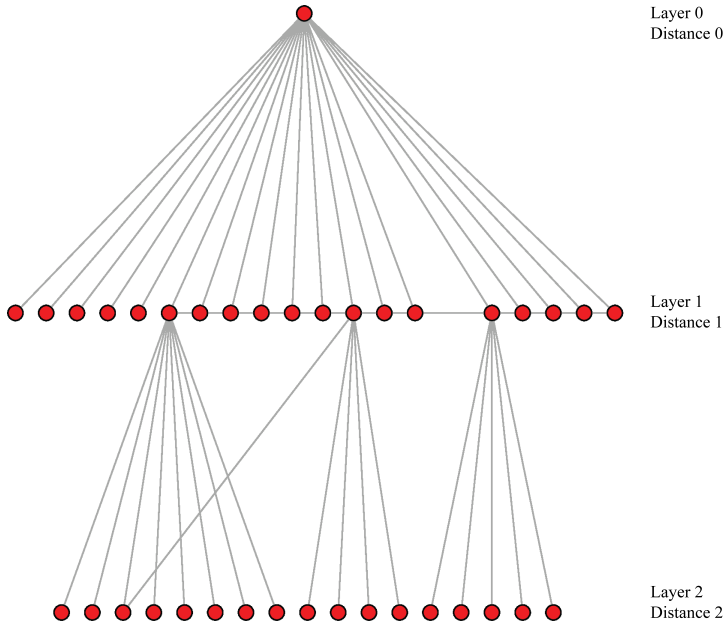


Figure 2.3: Example of a Breadth-first Search starting at my personal Twitter account. Layer 0 is the root at a distance of 0. Layer 1 represents root Followers at a distance of 1. Layer 2 represents Followers of three arbitrary root Followers at a distance of 2.

From these experiments we can perceive that this type of analysis demands a huge computational effort, since one is dealing with hundreds of millions of users as in the Microsoft Instant Messaging case [56], but also other problems must be considered, such as educational and technological background (people tracked in the experiment need to be technologically endowed enough to use an Instant Messaging service). In fact, even when all these problems are considered, it can be hard to understand the results. For instance, in the six degrees of separation experiment it does not necessarily mean that the majority of the population is socially "close".

There are other similar studies regarding network distances in smaller networks, such as the world of science. A common example, in Mathematics, is the *Erdös Number*, which gives the distance between some researcher and Paul Erdös, who was a prolific author having collaborated in a huge number of publications in Mathematics, and considering that there is a relation between two researchers when they publish a paper together. In the world of cinema, there is the *Bacon Number* that gives the distance between two actors which considers that there is a relation between two actors when they participated in at least one movie together. An experiment using The Internet Movie Database – website (IMDb) [1] obtained an average Bacon Number of 2.9 [24].

With the increasing availability of large and detailed network datasets, research on large-scale networks has increased massively. However, the datasets must have characteristics that allow its study, such as containing structured data. There are three major reasons to study a particular dataset: 1) The interest in the actual domain of the dataset, so that fine-grained details are potentially as interesting as the broad picture; 2) The availability and need for using a dataset as a proxy for a similar network that is impossible to measure; 3) The interest in looking for network properties that seem to be common across many different domains, which could even be universal. Of course, the three reasons can be present simultaneously in the same research.

[1] `http://www.imdb.com/`

2.1.2 Centrality Measures

The importance and prominence of a node depends on its position in the network. *Centrality measures* are properties that quantify the importance value of a node n in a graph G, and the most interesting ones that one can define are *Degree Centrality*, *Closeness Centrality*, *Betweeness Centrality* [104].

In order to be able to define these global properties that relate n with its network, it is important to start by analyzing its local relationships with its neighbor nodes. Node degree counts the number of those relationships.

Definition 7 (Degree) *$deg_G(n)$ is the sum of all nodes to whom n is connected. If k is the number of nodes in the network, the degree of n is at most $k-1$.*

For undirected networks, considering $(n,m) = (m,n)$ as the same edge between node n and node m belonging to the set of all edges that include n in G, degree is defined as:

$$deg_G(n) = \sum_m (n,m) = \sum_m (m,n) \tag{2.3}$$

For directed networks, the degree of n is measured according to the two different directions separately, because (n,m) and (m,n) represent two different edges between n and m. Considering the number of incoming edges, pointing to n,*in-degree* is defined as:

$$deg_G(n)^{in} = \sum_m (m,n), (n,m) \neq (m,n) \tag{2.4}$$

Considering the number of outgoing edges, pointing from n, *out-degree* is defined as:

$$deg_G(n)^{out} = \sum_m (n,m), (n,m) \neq (m,n) \tag{2.5}$$

By combining these individual properties with general information about the network, it is possible to define *Degree Centrality*.

Definition 8 (Degree Centrality) *$D_G(n)$ is the fraction between node n degree and the number of all k nodes of G.*

$$D_G(n) = \frac{deg_G(n)}{k-1} \tag{2.6}$$

This definition can also be applied to the graph G itself, considering $n*$ as the node with highest degree centrality and N as the set of nodes in G:

$$D_G(G) = \frac{\sum_{i=1}^{|N|} [D_G(n*) - D_G(n_i)]}{n-2} \tag{2.7}$$

Related with the distance from n to the rest of the network one can define Closeness Centrality.

Definition 9 (Closeness Centrality) *$C_G(n)$ is the inverse of the sum of distances between n and each node m of G reachable from n.*

$$C_G(n) = \frac{1}{\sum_m d_G(n,m)} \tag{2.8}$$

Alongside with the distance, the node's position can also be a bridge between several other nodes in G. This notion is measured by Betweeness Centrality.

Definition 10 (Betweeness Centrality) *$B_G(n)$ is the fraction between the sum of the number of shortest paths between all pairs (m_1, m_2) of nodes in N of G that include n – $s_{m_1 m_2}(n)$, and the total of shortest paths between that same nodes – $s_{m_1 m_2}$.*

$$B_G(n) = \sum_{m_1 \neq m_2 \neq n \in N} \frac{s_{m_1 m_2}(n)}{s_{m_1 m_2}} \tag{2.9}$$

Regarding these measures, a node n can be classified as:

Definition 11 (Pivotal node) *n is pivotal for a pair of distinct nodes m_1 and m_2 if n lies on every shortest path between m_1 and m_2, and they are all different nodes.*

$$s_{m_1 m_2}(n) = s_{m_1 m_2}, m_1 \neq m_2 \neq n \in N \tag{2.10}$$

Definition 12 (Gatekeeper) *n is gatekeeper for a pair of distinct nodes m_1 and m_2 if every path from m_1 to m_2 passes through n, and they are all different nodes.*

2.1.3 Tie Strength and Network's Dynamic

Networks not only represent large-scale relations on their subsets, but they also act as a bridge between local relations and global components. Explanations for how simple behaviors from individual nodes and edges can ripple through a population as a whole can be found on those relations and they can also explain why they possibly turn into complex effects.

Focusing on Social Networks, it is important to define some characteristics that can help understanding how information flows, how different nodes play structurally distinct roles, how these characteristics shape the evolution of the network itself overtime.

Two important properties are *Triadic Closure* and *Clustering Coefficient*:

Definition 13 (Triadic Closure) *If two persons in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.*

Definition 14 (Clustering Coefficient) *$CC_G(n)$ is the probability that two randomly selected friends of n are friends with each other. Assuming $P_{N_G}(n)$ as the number of pairs of neighbors connected to n that are also connected among themselves in graph G, and $N_G(n)$ as the total number of neighbors connected to n, clustering coefficient is defined as:*

$$CC_G(n) = \frac{P_{N_G}(n)}{\binom{N_G(n)}{2}} \tag{2.11}$$

Figure 2.4 shows the clustering coefficient for node with label 1.

The triadic closure is an intuitive property for which we can easily find examples in our personal life, and it is illustrated in Figure 2.5. *Opportunity to meet*, *trust*, and *incentive* are considered the three main reasons for the phenomenon to happen. Since n spends time with m_1 and m_2, there is an increased chance of m_1 and m_2 to have an opportunity of meeting each other. Person n will also be a trust factor between m_1 and m_2, and eventually, n can directly incentive m_1 and m_2 to know each other.

Bearman and Moody [11] used these properties to find that teenage girls with a low clustering coefficient in their network of friends are significantly more likely to commit suicide than those whose clustering coefficient is high.

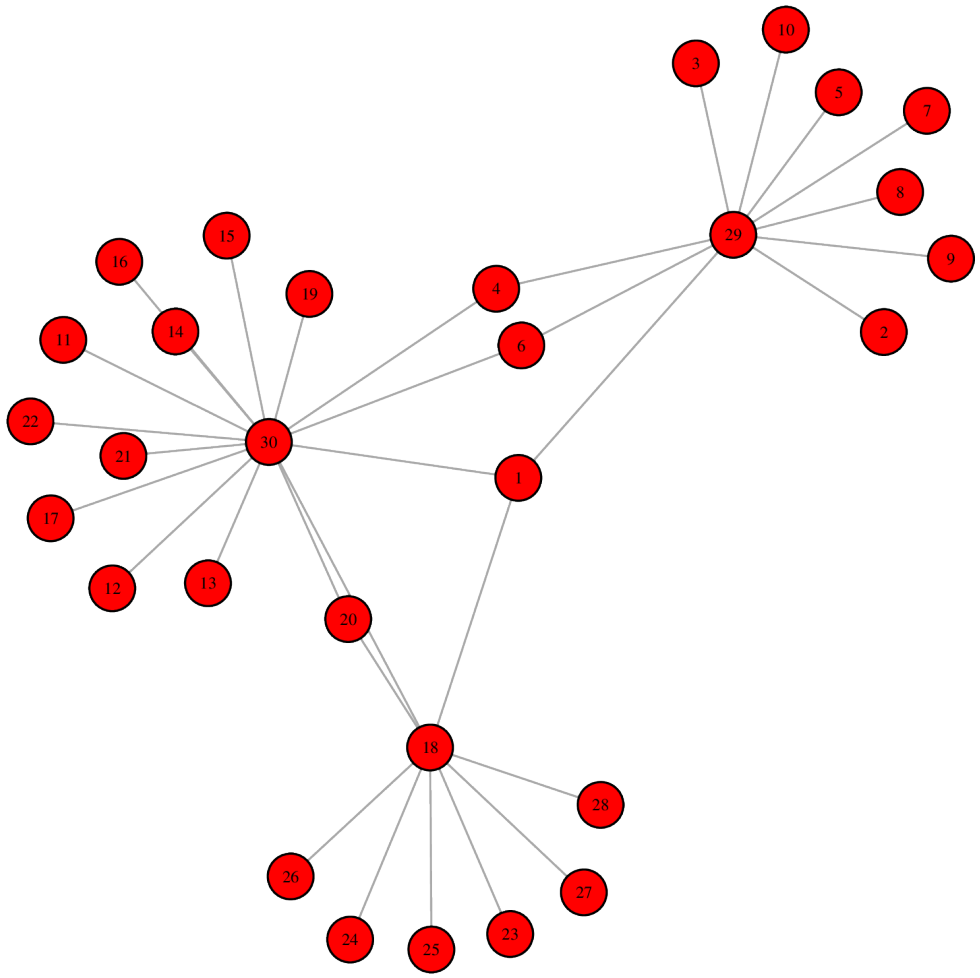


Figure 2.4: Node 1 has three neighbors: 18, 29, 30. From the three possible pairs of neighbors, only the pair (18, 30) is also connected, which gives $CC_G(1) = \frac{1}{3}$.

Regarding the connection between different size components such as local networks and global ones, there is the notion of ties' strength (related to the links that connect them). Ties can either be Weak or Strong, for example in social networks, edges between *friends* are *strong ties* and edges between *acquaintances* are *weak ties*.

2.1.4 The Leading Role of Weak Ties

Weak Ties actually have a dual role. They are weak but also valuable because they connect hard-to-reach parts of the network, playing a major role between connected components.

Definition 15 (Bridge) *There is a* bridge *between n and m when there is an edge between them and, in case of deleting this edge, n and m would be separated in two different components without a connection.*

However, bridges are extremely rare in real social networks and configurations like Figure 2.6 are unusually seen. In most cases, even friends with very different backgrounds have other paths to get to each other, they are just not aware of them [24]. Therefore, there is a refined definition that is called *Local Bridge*, which is represented in Figure 2.7.

Definition 16 (Local Bridge) *An edge between n and m is a local bridge if they do not have any friend*

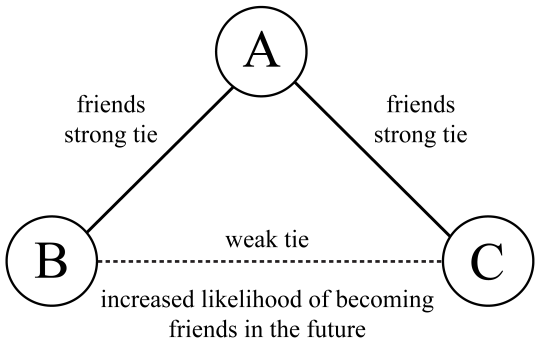


Figure 2.5: When two individuals have a common friend, they are probably aware of each other and there is an increased chance of being friends in the future.

in common, which means that deleting this edge will increase their distance to a value strictly more than 2.

Definition 17 (Span of Local Bridge) Span *of a local bridge* $S_G(e), e = (n, m) \in E$ *is the distance between* n *and* m*, if the local bridge between them were deleted,*

$$S_G(e) = d_{G \setminus e}(n, m) \tag{2.12}$$

Combining the concepts of Triadic Closure with Strength in links of a social network, entails the notion of *Strong Triadic Closure*.

Definition 18 (Strong Triadic Closure) *Some node* n *violates the strong triadic closure property if it has strong ties to two other nodes* m_1 *and* m_2*, and there is no edge at all (strong or weak) between* m_1 *and* m_2*. Otherwise* n *satisfies the strong triadic closure. Also, if a node* n *in a network satisfies the strong triadic closure property and it is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.*

Given these theoretical concepts, it is important to understand how to apply them with real data on large scale. This could be done defining numerical quantities and sometimes it could be also useful to smooth the definition. For instance, with a dataset of phone calls, the strength of ties could be given by the number of minutes spent on phone calls. A way to find local bridges is getting the *Neighborhood Overlap* $O_{N_G}(e)$. Assuming that $e = (n, m) \in E_G$, and $N_{n \cap m}(n, m)$ is the number of neighbors of both n and m, and $N_{n \cup m}(n, m)$ is the number of nodes that are neighbors of at least on n or m, then

$$O_{N_G}(e) = \frac{N_{n \cap m}(n, m)}{N_{n \cup m}(n, m)} \tag{2.13}$$

An edge is a local bridge if the neighborhood overlap is 0. Depending on the sample and the purpose, the occurrence of local bridges could be low (or inexistent) and smaller values of neighborhood overlap could be accepted as “almost” local bridges.

Nowadays, a huge amount of social interactions occurs online and this is both creating a new paradigm and also changing the way that people maintain, perceive and access their social networks. A familiar phenomenon in social networks like Facebook and Twitter is that people maintain a large list of friends in their profiles, much larger than they would enumerate mentally without a record. People can

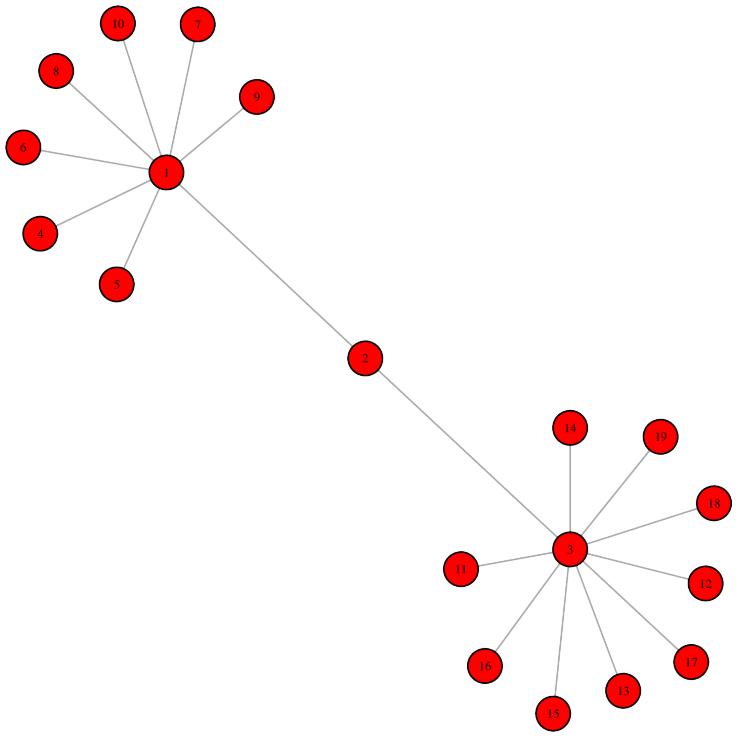


Figure 2.6: Every edge in this figure is a bridge. Whichever edge that could be removed, it separates its endpoints into two independent components.

have hundreds of links to other people, but they may not represent strong ties. The frequency of contact between those links can help in finding their true strength.

2.1.5 Power and Place in the Network

The position of a node, regarding closure properties, can influence its power inside the network. Understanding this property can help us in managing social capital.

Definition 19 (Embeddedness) *Embeddedness of an edge in a network is the number of common neighbors shared by the two endpoints,* $N_G(e), e = (n, m) \in E_G$*. The higher the embeddedness, the higher trust, confidence in integrity and reputation between them.*

The formal definition of local bridges shows the importance and the power of certain node positions in a network, in a mathematical way. One node at the end of multiple local bridges aggregates a distinct set of equally fundamental advantages and it is defined as a *Structural Hole*. This is an informal definition regarding the qualitative importance of its position. Some possible advantages are: early access to multiple kinds of information, contacts, amplifier of creativity, social "gatekeeping". *Social capital* is a term used in sociology about individuals and groups deriving benefits from an underlying social structure or network, where physical capital refers to material goods that bring value to something (like technology), human capital refers to the individual talents and skills of people, economic capital refers to monetary and physical resources, and cultural capital refers to the accumulated resources of a culture beyond individual social circles (such as education level). Social capital stands as a framework for thinking about social structures as facilitators of effective action by individuals and groups, combining specific characteristics of the individuals with the properties of their underlying network.

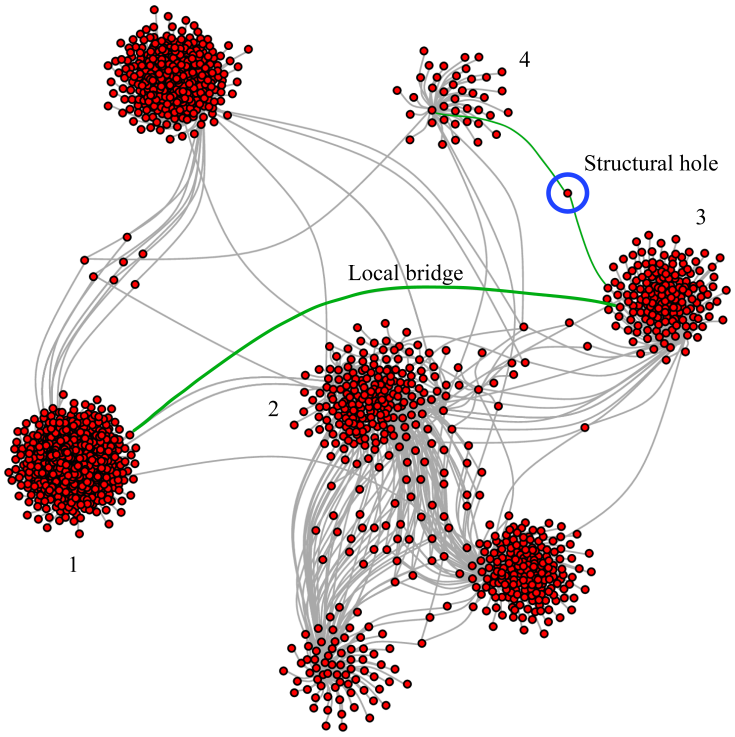


Figure 2.7: The local bridge highlighted in green decreases the distance between the two densely connected components to which it is attached (1 and 3). Without this edge, the path between these two components would cross another component (2). The node highlighted in blue is a structural hole because it is connected to two densely connected components (3 and 4), through two local bridges.

2.1.6 Popularity Models

Some phenomenons are intrinsically related with the Dynamics of Social Networks, such as the previously discussed concept of Influence. Related to this, there is also the notion of *Popularity*.

Popularity is characterized by extreme imbalances because, while almost everyone goes through life known only to people in their immediate social circles, a few people achieve wider visibility, and a very, very few attain global name recognition. The Web is an interesting network to think about popularity because web search engines work taking into account web pages' popularity. The popularity of a web site is defined as the number of existing links pointing to it, named *in-links* [24].

Thinking about its distribution regarding the Web example, if we take different Web snapshots, at different points in time, the fraction of web pages that have k in-links is approximately proportional to $1/k^2$ (more precisely, the exponent on k is generally a number slightly larger than 2) [24]. This distribution follows the *power-law form*.

Definition 20 (Power-law) *There is a power-law relation when the fraction of items k is a function:*

$$f(k) = \frac{a}{k^c}, \tag{2.14}$$

for some exponent c and constant of proportionality a.

A possible approach to model the dynamics of popularity for a given set of nodes starts by associating the first links randomly among them. Then, the following links are created choosing a link uniformly at random from the earlier links, and the new link will point to the same node that the previously created link points to. The randomness makes hard for some nodes to become popular, but as long as the links that point to them are copied, the probability of being chosen increases according to the power-law

[24]. These models combine randomness with information cascades. Although cascades explain why *rich-get-richer*, the initial random assignments do not show how popular nodes begin to be chosen in the beginning. The initial dynamics responsible for book sales, songs' success, technological devices sales, is probably not properly modeled by randomness.

2.1.7 Relationship Polarity and Network's Shape

So far we have considered different network relationships varying from friendship, collaboration, sharing of information, membership in a group. However, there are also antagonistic or hostile relations such as controversy, disagreement, and conflict.

These types of relationships can be grouped as *positive* or *negative* relationships [104]. Relations in a network are represented by edges and, in order to classify them as positive or negative, these edges must be labeled. For example, edges annotated with (+) may represent friendship and (-) may represent antagonism [80]. This representation is useful to study tension between these two forces, understanding their evolution and mutation over time.

International relations and political science are two research areas with great interest in capturing mathematical properties and rules from networks with polarized relationships. By finding them, effective explanations for the behavior of nations and their positions among each other, during international crisis for example, could be provided [24].

Structural Balance and The Balance Theorem are two properties that evolve from local phenomenons to global consequences in dynamic polarized networks [80].

Definition 21 (Structural Balance) *for every set of 3 nodes, either all 3 are labeled positive (+), or exactly 1 is labeled positive (+).*

Definition 22 (The Balance Theorem) *If a labeled complete graph is balanced, then either all pairs of nodes are friends, or the nodes can be divided into 2 groups, X and Y, such that each pair of people in X likes each other, the same in Y, but everyone in X is enemy of Y and vice versa.*

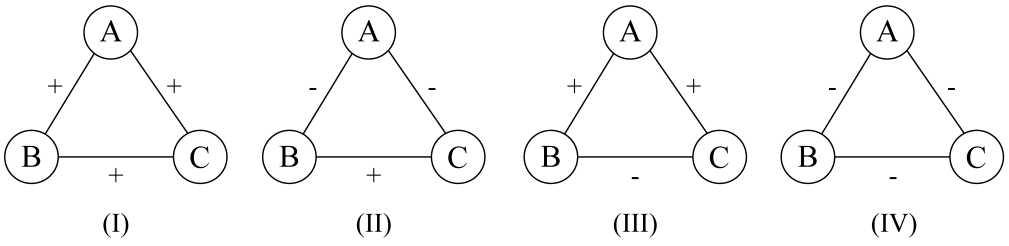


Figure 2.8: Illustration of Structural Balance [82], (I) and (II) are balanced, while (III) and (IV) are not.

The global effect of Structural Balance is verified in a network if it respects the *Weak Structural Balance property*.

Definition 23 (Weak Structural Balance property) *There is no set of 3 nodes such that the edges among them consist of exactly 2 positive edges and 1 negative.*

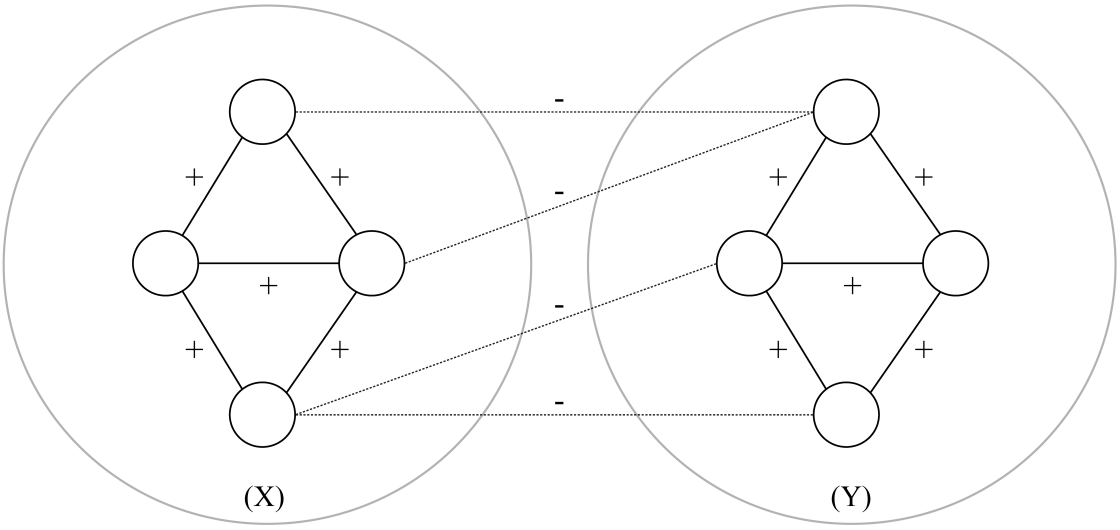


Figure 2.9: A balanced graph with both positive and negative relations can be divided in two groups, (X) and (Y), in such away that all the edges inside each group are positive and all the edges between the two groups are negative, according to the Balance Theorem.

If a labeled graph is weakly balanced, then its nodes can be divided into groups in such a way that every two nodes belonging to the same group are friends and every two nodes belonging to different groups are enemies, respecting the Balance Theorem. This is illustrated in Figure 2.9.

A signed graph is balanced if and only if it contains no cycles within an odd number of negative edges. Once again, this is not a static property and the labeling of the edges changes over the time. In practice, the available graphs are not necessarily complete nor truly balanced. Research in this topic tries to understand in which way unbalanced graphs evolve into balanced or "approximately balanced" graphs.

Antal, Krapivsky and Redner [4] demonstrated that shifting alliances preceding World War I evolved according to structural balance and that escalated to a global conflict between two opposite alliances. Online ratings are also a source of data for networks with positive and negative edges, but instead of representing friendship and antagonism, these labels represent trust and distrust, respectively. Guha et al. [39] performed an analysis on a network of user evaluations on Epinions [2] and they identified both similarities and differences between trust-distrust dichotomy and friend-enemy dichotomy. A subtle difference is related to the structural balance, and the question is: when A distrusts B and B distrusts C, does A trust or distrust C? They found that the answer is not trivial and depends on the topic. For example, in rating political books, it is expected that A trusts C, because A is not politically aligned with B and neither B with C, so it is expected that A and C have close political orientations. On the contrary, in rating electronic products, the tendency is for A to distrust C because A distrusts B considering himself more expert on the topic, and since B distrusts C, B should be more expert than C, for what A assumes he is also more expert than C.

Both Structural Balance and the Balance Theorem are based on undirected relationship networks, like friendship networks, but their evidence is not observed in directed relationship networks [57]. In directed networks, like a network of opinions about colleagues' skills among the players of a soccer team, Social Status Theory is observed. In a directed graph, each directed relationship labeled positive (+) denotes the target node has higher status than the source node, and each directed relationship

[2] http://www.epinions.com/

labeled negative (-) denotes the target node has lower status than the source node.

Definition 24 (Social Status Theory) *For every set of 3 connected nodes that forms a triad, we take each negative edge, reverse its direction and flip its sign to positive, then the resulting triangle (with all positive edge signs) should be acyclic [82].*

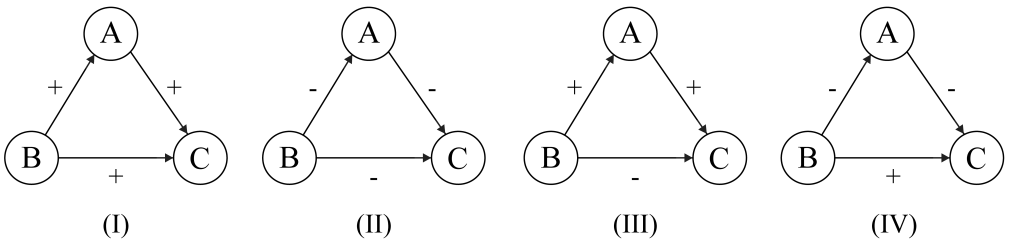


Figure 2.10: Illustration of Social Status Theory [82]. (I) and (II) satisfy Status Theory, while (III) and (IV) do not.

2.1.8 Social Similarity and Context Surrounding Influence

Looking at our set of friends, they are generally similar to us and among them in some characteristics like age, ethnicity, place where they live, among others. This similarity also exists in other characteristics more or less mutable, like occupation, interests, beliefs, opinions. Disparities in some of these characteristics exist and some of our friends can cross all these boundaries, but in aggregate, links in a social network tend to connect people who are similar to one another, and that is known as *Homophily*.

Definition 25 (Homophily) *Homophily is the principle that states that we tend to be similar to our friends.*

This principle can divide a social network into densely connected, homogeneous parts that are weakly connected to each other. In order to understand the expression of Homophily, for a certain feature of interest in a given social network, links between nodes with opposite features must be counted. The proof for Homophily is verified by refuting the no-homophily scenario.

Definition 26 (Homophily Test) *If the fraction of edges between nodes with opposite features $f1$ and $f2$ is significantly less than $2 * P_f1 * P_f2$, where P_f1 and P_f2 are the probability of choosing a node with feature $f1$ and $f2$, respectively, then there is evidence for Homophily. The fraction's boundary values that distinguish between homophily and no-homophily must be defined according to the test's purpose.*

Let us consider the example of trying to figure out if there is gender homophily in a friendship network of an elementary-school class. We know that P_m is the probability of a node to be male and P_f is the probability of a node to be female. No-homophily exists with the prevalence of cross-gender edges. The probability of a cross-gender edge is $2 * P_m * P_f$ (considering the case that the first node is male and the second female, and vice versa), if the fraction of cross-gender edges is significantly less than the probability of a cross-gender edge, there is evidence for Homophily.

The existence of Homophily is consequence of both *Social Selection* and *Social Influence* [24].

Definition 27 (Social Selection) *Selection reflects the tendency of people to form friendships with others who are like them, according to a set of similar characteristics.*

Definition 28 (Social Influence) *Social Influence is the tendency of people, integrated in a social circle, to adopt some shared characteristics of people in that circle.*

Social Influence is commonly associated to peer pressure, and can be viewed as the reverse of Selection, since in selection the individual characteristics drive the formation of links, while with Social Influence, existing links in the network serve to shape people's (mutable) characteristics. One of the major causes of delinquent behavior and drug use is believed to be peer pressure [24]. But this analysis needs to consider many variables and distinguish among different factors which can lead to unclear conclusions.

All these considerations relate contextual factors with the formation of links in a network, based on similar characteristics and common behaviors or activities, where surrounding contexts were viewed as existing outside the network. However, to properly analyze the evolution of friendships together with their context changes, this set of activities could be put on a similar network, called *Affiliation Network*.

Affiliation Networks are bipartite graphs with a set of nodes representing people nodes connected to nodes in the set of nodes *foci*. Foci are focal points of social interaction that represent activities or aggregation of activities (such as workplaces, voluntary organizations, meetings). In this way, people are explicitly and directly associated with their characteristics, which could be shared among other people.

Both Social and Affiliation Networks change over time, since friendship links change, end or new ones are formed, at the same time that people change associations to their foci, end and create new ones. Analyzing this coevolution helps inferring their mutual influence. Analyzing their evolution also can be used to prove the Triadic Closure property and answer deeper questions related to it, like *focal closure – the probability that two people form a link as a function of the number of foci they are jointly affiliated with* – and *membership closure – the probability that a person becomes involved with a particular foci as a function of the number of friends who are already involved in it*. With these two notions it is also possible to quantify the interplay between Selection and Social Influence.

Homophily is observed in practice, having ethnicity as an example of feature, as one cause of formation of homogeneous groups, that can scale up to neighborhoods and entire cities. Inherent to this analysis is the dynamic of these networks, regarding time and space. Link evolution can be observed in a sequence of rounds and can entail space changes [24].

2.1.9 Influence and Information Cascades

People connected to other people are susceptible to be influenced by them. People in a social network have the power to influence others connected to them in that network and the network can influence them too. Decisions and behaviors can be influenced, and people imitate others opinions, political views, choice of products, technologies, and so on.

Definition 29 (Decision Imitation) *Imitation happens when an individual, under influence, imitates the choices of others over, or not, his own information about the alternative choices [24].*

Although imitation may occur due to social pressure to conform, there are many situations in which following others decisions is in fact the most rational decision.

Supposing we are visiting an unfamiliar town and we want to choose a restaurant. Before arriving, we have searched some information about restaurants for instance in a guide and we read about restaurant A, but when we found restaurant A we discover that there is a restaurant B next door that is nearly full and A is empty of clients. In this case, we could consider that people eating at B have more powerful

information about these restaurants than our own private information. If we choose B, influenced by this inference, an information cascade has occurred [24].

Definition 30 (Information Cascades) *Information cascades happen when someone abandons his own information in favor of inferences based on earlier people's actions and are more likely to occur when these actions are sequential.*

According to the experiments of Milgram, Bickman and Berkowitz [67], the likelihood of imitating some common action of a group of people grows with the number of people in that group. In some cases, the reason for following others' decisions could be that aligning both behaviors brings direct benefits. An example of that is choosing a mobile operator.

In order to understand the logic behind information cascades, we will consider a simple herding experiment created by Anderson and Holt [2]. This experiment is based on the following assumptions:

1. There is a decision to be made.
2. People make the decision sequentially, and each person can observe the choices made by those who acted earlier.
3. Each person has some private information that helps guide their decision.
4. A person cannot directly observe the private information that other people have, but he can make inferences about it from their actions.

The experiment takes place in a classroom, with a large group of students as participants. There is an urn with three marbles hidden inside and there is a 50% chance that two of them are red and one is blue, and 50% chance that two of them are blue and one is red. The challenge to each student, one by one, is to draw a marble, without showing it to the rest of the class, and then publicly announce to the class his guess about whether the urn is majority-red or majority-blue. At the end of the experiment, each student who has guessed correctly receives a monetary reward, the others receive nothing. The key part of the experiment is the public announcement, that allows the next participant to infer what happened before.

Considering that all students make a rational decision, it is expected that the decisions would be:

- *First student:* there is no further information but the color of the marble taken, so he must guess only according to the only piece of information that he has seen only.
- *Second student:* If the color he sees is the same as announced by the first student, his choice should be this color as well. However, if the color is different he is indifferent about which guess to make. Again, what he sees is taken into account in every scenario.
- *Third student:* If the first two students have guessed different colors, he must guess based on what he sees and should break the tie. But considering the case where both earlier students have guessed the same color, the third student knows that this conveyed perfect information about what they saw, so he must state the same color ignoring his own private information. In this case, an information cascade has begun.
- *Fourth student and onward:* if all the first three students have guessed the same color, all the remaining students will be in the same situation as the third student, and the cascade will continue.

From this example we can perceive that information cascades can persist for a long time based on poor and wrong information. If in the case of a cascade that has begun, as in the previous experiment,

some students cheated and have shown the color they have seen, this could end the cascade because they would give more precise information to the next students. That is why information cascades can be very fragile.

Building mathematical models about information cascades implies determining the probabilities of events given the information that is observed. For that, one uses conditional probabilities and Bayes' theorem.

Assuming that $P(A)$ and $P(B)$ are the probabilities of the events A and B to occur, respectively, and that $P(A \cap B)$ is the probability of both to happen. The conditional probability of an event A to occur, knowing that event B happened is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.15}$$

Knowing that:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \tag{2.16}$$

This leads to the Bayes' Theorem:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} \tag{2.17}$$

Applying this to the first student in the herding experience we can say that the probability of the urn to be majority-blue (MB), knowing that the he draws a blue marble (B) is:

$$P(MB|B) = \frac{P(MB) * P(B|MB)}{P(B)} \tag{2.18}$$

The denominator is the most tricky variable because, like in the example, it represents the probability of the marble to be blue considering that both scenarios, that the urn is majority-blue or majority-red, are possible.

We calculate that with the Law of Total Probabilities, assuming that $A_1, ..., A_n$ are the different possible scenarios:

$$P(B) = \sum_{i=1}^{n} P(B|A_i) * P(A_i), \forall B \in A \tag{2.19}$$

But it is important to bear in mind (1) the possible states that are being analyzed (like the first student without any further information and the remaining students that knew the previous guesses), (2) the possible payoffs resulting from the possible decisions and (3) the private signals related to the private information.

To sum up, let's consider the perspective of person N:

- If the number of acceptances among the people before N is equal to the number of rejections, then $N's$ will be the tie-breaker, and so N will follow her own signal.
- If the number of acceptances among the people before N differs from the number of rejections by one, then either N's private signal will make her indifferent or it will reinforce the majority signal. Either way, N will follow her private signal (we assume a person follows their own signal in the case of indifference).
- If the number of acceptances among the people before N differs from the number of rejections by 2 or more, then however $N's$ private signal turns out, it will not outweigh this earlier majority. As a result, N will follow the earlier majority and ignore her own signal. Moreover, in this case, the people numbered $N + 1$, $N + 2$, and onward will know that person N ignored her own signal (whereas

we have assumed that all earlier people were known to have followed their private signals). So they will each be in exactly the same position as N. This means that each of them will also ignore their own signals and follow the majority. Hence, a cascade has begun.

Although cascades may be wrong, when based on very little information, and fragile, their effects can persist in time and have powerful consequences. Understanding this phenomena is of great interest in many areas, such as product sales, stock market, political elections.

2.1.10 Influence and Cascading Behavior

As it was said before, there are two distinct kinds of reasons why imitating the behavior of others can be beneficial: informational effects, based on the fact that the choices made by others can provide indirect information about what they know; and direct-benefit effects, in which there are direct payoffs from copying the decisions of others.

Both are distinct but both can be connected, such as in diffusion of innovations. For example, in the case of the adoption of a new operating system, despite the fact that it may bring relative advantage, compared with the existing solutions, its success depends on its *complexity* for people to understand and implement; *observability*, so that people can become aware that others are using it; its *trialability*, so that people can mitigate its risks by adopting it gradually and incrementally; and, perhaps most crucially, its overall *compatibility* among peers in that network.

Related to these issues, the principle of Homophily can sometimes act as a barrier to diffusion since people tend to interact with others who are like themselves, while new innovations tend to arrive from "outside" the network. Therefore, it can be difficult for innovations to make their way into a tightly-knit social community.

Network models based on direct-benefit effects consider the idea that the benefits to you of adopting a new behavior increase as more of your social network neighbors adopt it. Assuming that you will make a decision by self-interest, you should adopt the new behavior once a sufficient proportion of your neighbors have done so.

This idea can be captured by using a coordination game. In an underlying social network, each node v and w has a choice between two possible behaviors, labeled A and B. If nodes v and w are linked by an edge, then there is an incentive for them to have matching behaviors. The possible strategies can be represented by a payoff matrix considering that:

- if v and w both adopt behavior A, they each get a payoff of $a > 0$;
- if they both adopt B, they each get a payoff of $b > 0$;
- if they adopt opposite behaviors, they each get a payoff of 0.

		w	
		A	B
v	A	a, a	0, 0
	B	0, 0	b, b

Table 2.1: Payoff matrix of *w* and *v* choosing behavior A or B

There are two equilibriums in this network-wide coordination game: one in which everyone adopts A, and another in which everyone adopts B. But, in the initial or intermediate stages, both can coexist, where A is adopted in some parts of the network and B is adopted in others.

A simple threshold rule can explain when a person should adopt, for example, behavior A: it says that if a fraction of at least $q = b/(a + b)$ of your neighbors follows behavior A, then A is more interesting

for you. In the case when there already is a dominant behavior, let us assume B, it can happen that behavior A enters in the network by being adopted by initial adopters. In such case, A could become dominant if the number of initial adopters and the payoff values of a and b becomes more valuable for them.

When this chain reaction of switches to A occurs, there are two distinct possibilities: a cascade runs for a while but stops while there are still nodes using B; or there is a complete cascade and every node in the network switches to A.

Considering a set of initial adopters who start with a new behavior A, while every other node starts with behavior B. Then, each node repeatedly evaluates the possibility of switching from B to A using a threshold of q. If the resulting cascade of adoptions of A eventually causes every node to switch from B to A, then we say that the set of initial adopters causes a complete cascade at threshold q.

Homophily can often serve as a barrier to diffusion by making it hard for innovations to enter in densely connected communities.

We say that a cluster of density p is a set of nodes such that each node in the set has at least a fraction p of its network neighbors in the set.

Considering a set of initial adopters of behavior A and threshold q for nodes in the remaining network to adopt behavior A:

- If the remaining network contains a cluster of density greater than $1 - q$, then the set of initial adopters will not cause a complete cascade.

- Moreover, whenever a set of initial adopters does not cause a complete cascade with threshold q, the remaining network must contain a cluster of density greater than $1 - q$.

It was found that clusters are the only obstacles to cascades [24]. Whenever a set of initial adopters fails to cause a complete cascade with threshold q, there is a cluster in the remaining network of density greater than $(1 - q)$. From this it can be said that a set of initial adopters can cause a complete cascade with threshold q if and only if the remaining network contains no clusters of density greater than $1 - q$. Clusters block the propagation of cascades and whenever a cascade breaks, it is due to a cluster with a density higher than $1 - q$.

Weak ties have shown that learning about a new idea is crucially different from actually deciding to adopt it, in a social network. As it was already seen, local bridges are important sources of information between different groups in a network – for instance, announcements of new job opportunities. Sometimes, local bridges can be the only form by which certain information can reach a certain group. However, to propagate a new behavior it requires not just awareness, but also a threshold value for that group to adopt it.

Nonetheless, there are situations in which all we have described so far is not enough, since it also requires *coordination* across a large segment of the population. A characteristic problem of collective actions is when the activity only produces benefits if enough people participate, and people do not know their peers' *willingness*. To overcome this problem one has to find a way to transmit information about *people's willingness* through the social network.

An illustrative case of this problem is organizing a protest or revolt under a repressive political regime. It may be the case that, in spite of there being a common will among a large number of people in the network, that will is not shared among them due to fear of reprisals. This lack of knowledge about others' will can create an erroneous estimate about the prevalence of certain options in the population at large. This phenomenon is known as *pluralistic ignorance*. In fact, a massive protest or revolt can weaken or even put down a government, but if a large number of people stay at home, the small group that decides to show up can suffer severe consequences. From this, it can be modeled that each person, knowing

about a potential upcoming protest, has a personal threshold k as a minimum number of participants (including him) to participate too. This threshold also encodes that person willingness to participate.

Strong ties have an important role in coordinating and spreading knowledge about people's willingness through the network, because people strongly connected tend to be similar, and to have overlapping knowledge. Once coordination is needed, weak ties are not enough to allow information propagation, instead it requires social institutions widely connected to the different components of the network, such as a widely publicized speech, or an article in a high-circulation newspaper, building a common knowledge among weakly connected components. They not only ensure the transmission of the message, but also make the listeners or readers realize that many others have received that message too.

People not only base their decisions on what others know, but also on how they expect others to behave as result.

The *cascade capacity* of a network is the largest value of the threshold q for which some finite set of early adopters can cause a complete cascade. Easley and Kleinberg [24] argue that there is no network in which the cascade capacity exceeds $\frac{1}{2}$.

2.1.11 Information Diffusion and Epidemics

The study of epidemic diseases mixes biological issues with social ones. These are contagious diseases caused by biological pathogen agents which propagate from person to person, such as influenza or Sexual Transmitted Diseasess (STDs). Epidemics can propagate explosively through a population, or they can persist over a long time period at low levels; they can also experience sudden flare-ups or even wavelike cyclic patterns of increasing and decreasing prevalence. In extreme cases, some epidemics can have major effects on a whole civilization.

The patterns by which epidemics propagate through groups of people is determined not just by the properties of the pathogen agent carrying it (contagiousness, length of infectious period, severity), but also by the network structures within the populations it is affecting. The opportunities for a disease to propagate are given by a contact network, in which each node stands for each person, and an edge represents whether two people come into contact with each other in a way that makes it possible for the disease to propagate from one to the other.

There are clear connections between epidemic diseases and diffusion of ideas through social networks, both propagate from person to person across similar kinds of networks that connect people. This is the reason why spreading ideas is called *"social contagion"*. The major difference between them relies in the process of contagion, since the diffusion of ideas can imply a decision-making process and, on the other hand, a transmission of a disease may not. Random is a common approach to model the transmission of a disease. A more refined model about epidemics may consider synchronization, timing, and concurrency in transmission.

A simple model of contagion, known as branching process works as follows:

- *First wave:* Suppose that a person, who we call root, carrying a new disease, enters a population and transmits it to each person he meets independently with probability p. Furthermore, suppose that he meets k people while he is contagious; let us call these k people the first wave of the epidemic. Based on the random transmission of the disease from the initial person, some of the people in the first wave may get infected with the disease, while others may not.

- *Second wave:* Now, each person in the first wave goes out into the population and meets k new different people, resulting in a second wave of $k * k = k^2$ people. Each infected person in the first wave passes the disease independently to each of the k second-wave people they meet, again independently with probability p.

- *Subsequent waves:* Further waves are formed in the same way, by having each person in the current wave meet k new people and pass the disease to each independently with probability p.

There are two possibilities for a disease in the branching process model: it reaches a wave where it infects no one, thus dying out after a finite number of steps, or it continues to infect new people in every wave, proceeding infinitely through the contact network. These two possibilities depend on the *basic reproductive number* of the disease.

The basic reproductive number, denoted R_0, is the expected number of new cases of the disease caused by a single individual. Since everyone in our model meets k new people and infects each with probability p, the basic reproductive number, in such case, is $R_0 = pk$. The outcome of the disease in a branching process model is determined by whether the basic reproductive number is smaller or larger than 1.

If $R_0 < 1$, then with probability 1, the disease dies out after a finite number of waves. If $R_0 > 1$, then with probability greater than 0 the disease persists by infecting at least on person in each wave.

The Susceptible, Infectious, Removed(SIR) Epidemic Model [24] can be applied to any network structure and it states that an individual node in the branching process model goes through three potential stages during the course of the epidemic:

- *Susceptible (S)*. Before the node has caught the disease, it is susceptible to infection from its neighbors.

- *Infectious (I)*. Once the node has caught the disease, it is infectious and has some probability of infecting each of its susceptible neighbors.

- *Removed (R)*. After a particular node has experienced the full infectious period, this node is removed from consideration, since it is no longer a threat to future infection.

The progress of an epidemic is controlled by two additional quantities: p (the probability of contagion) and t_l (the length of the infection). Initially, some nodes are in the I state and all others are in the S state. Each node v that enters the I state remain infectious for a fixed number of steps t_l. During each of these t_l steps, node v has a probability p of passing the disease to each of its susceptible neighbors. After t_l steps, node v is no longer infectious or susceptible to further bouts of the disease. This node is removed because it becomes immune and no longer transmits the disease.

A variant of this model is the Susceptible, Infectious, Susceptible (SIS) Epidemic Model [24] for diseases that people don't become immune after recovering, in which the last step, instead of removing the node previously infected, returns it to the S state.

2.2 Twitter: A Wide Social Environment

Now that the standard concepts about Social Networks have been introduced, we focus state-of-the-art work on this are. The majority of the examples presented here use Twitter as their research basis, and are organized according to the following topics: *Networks Structure and Finding Communities*; *Event Detection*; *Event Prediction*; *Information Flow*; *Influence and Homophily*; *Sentiment Analysis: Positivity, Negativity, Neutrality*; *Spam Filtering*; and *Geo-location*. For each topic we summarize the major goals, their problems, their solutions and some interesting results.

For a better understanding of these studies it is important to define the main characteristic features of Twitter. As a Social Network service, Twitter connects registered users among them in a short message sharing environment.

Definition 31 (Tweeter User) *A user has a name and a unique username. Users can either be individual persons, as well as collective institutions such as companies, brands, organizations. There is no age or gender associated with the user. Users can add an image, description, language, location, and a timezone to their profiles.*

Users can follow an be followed by other users.

Definition 32 (Follow) *Following some user is integrating the list of users with whom that user will share his messages. Each user has a list of Followers – from whom he receives messages – and a list of Following users – to whom he shares his own messages. Users can protect their profiles and following them requires their approval.*

A message in Twitter is called Tweet. Posting a tweet is designated as twitting and it will appear in the user's profile, as well as in his followers's timelines.

Definition 33 (Tweet) *It is a message with a maximum size of 140 characters that can include photos and videos. An URL inside the message is always converted to a format of 22 characters and its content can also be displayed alongside with the message [101]. The location of the user when tweeting can also be added to the tweet.*

A tweet can be a simple text and/or media content, but it can also include direct interactions with other users. Retweeting or marking a tweet as favorite are content related actions, while mentions and replies represent conversational tweets. Hashtags can label a tweet as well as create topic related conversations.

Definition 34 (Retweet) *By retweeting a tweet, a user is forwarding that tweet to his own followers. It is commonly abbreviated as "RT". Each tweet has a counter of retweets.*

Definition 35 (Favorite) *This action saves the tweet in the user's list of favorite tweets. Each tweet has a counter of users that have marked it as favorite.*

Definition 36 (Mention) *Explicit reference to a user using the tag "@" followed by the unique username. For instance, typing "@hugomalopes" is a mention to the user "hugomalopes".*

Definition 37 (Reply) *A reply is a particular case of a mention, in which the mention is located at the bottom of the tweet. Replies are used to comment or answer to something that the mentioned user has tweeted.*

Definition 38 (Hashtag) *The "#" symbol, called a hashtag, is used to mark keywords or topics in a tweet. It was created organically by Twitter users as a way to categorize messages [102]. Clicking on a hashtagged word in any message shows all other available tweets marked with that keyword. Hashtags can occur anywhere in the tweet. The hashtag "#worlcup2014" can tag a tweet about the 2014 FIFA World Cup.*

Users interact with each other through Twitter web and mobile applications, but Twitter also provides two other programmatic channels to access its data: Representational State Transfer (REST) and Streaming Application Programming Interfaces (APIs). REST API provide programmatic access to read and write Twitter data and is used to conduct singular searches, read user profile information, or post Tweets. The Streaming API gives developers low latency access to Twitter's global stream of Tweet data, and it is used to monitor or process Tweets in real-time [99] [100].

2.2.1 Tie Strength on Twitter

Twitter allows different types of inter-user interactions, such as following, retweeting, mentioning. It is possible to infer different networks depending on the type of relation chosen to build the graph. However, it is broadly agreed that follows do not fairly represent real relationships among users [80, 16, 45]. Follow relations are bridges of information but do not represent direct communication between the users, acting as a passive behavior, characteristic of weak ties. On the other hand, retweets and mentions are interactions that represent direct communication, from which stronger relations can be inferred.

Huberman, Romero and Wu [45] analyzed the relative abundance of these two kinds of links on Twitter. For each user, they considered strong ties when he has mentioned another user at least twice, over the observation period. Then they compared the number of strong ties with the number of follows and even for users who follow more than 1000 other users (weak ties), the number of strong ties remains relatively modest, lower than 45.

Cameron Marlow and his colleagues [65] analyzed data of a one-month observation period of Facebook usage and they defined three (not mutually exclusive) categories of links: *Reciprocal (mutual) communication link:* when both users in the link sent messages between each other during that period; *One-way communication link:* when at least one user sent one or more messages to his friend in the link during that period; *Maintained relationship link:* when at least one user in the link follows information about his friend, by clicking on content via Facebook's News Feed service or visiting his profile more than once.

Beyond these observations, both studies of Marlow and his colleagues [65] and Bakshy et al. [10] claim that even without communicating with many of their friends, people continue reading their news, suggesting that even with few strong ties, there is still a connection between them. They say that media like Facebook enables this kind of *passive engagement*, such that events like a new baby born or engagement can propagate very quickly through this highly connected network.

Facebook and Twitter are environments with relative scarcity of strong ties, by contrast to the number of weak ties. This happens because strong ties require continuous investment of time and effort, while weak ties do not. [24]

2.2.2 Network Structure and Finding Communities

Twitter information is biased and noisy, which makes the process of representing the network and navigating through its relations a complicated task.

Finding communities and social circles with some common characteristic, such as sports, is very valuable for areas like marketing. Lim et al. [59] used profiles of celebrities in music, news, blogging and then follow their followers, in order to reach their networks.

According to Mcauley and Leskovec [66], Social Circles can be accurately detected using a combination of both network and profile information. Wu [110] observed the existence of a small fraction of population (0.05%) that is responsible for half of the posts with URLs in his dataset. This fraction of the population is composed by an elite of users with strong homophily, with celebrities following celebrities, media following media, bloggers following bloggers.

Social Circles are usually connected through weak ties. The method for inferring social ties across heterogeneous networks, followed by Tang and Kleinberg [82], uses Social Status Theory guiding supervised learning. According to their observations, most triads (99%) satisfy properties of Social Status and opinion leaders are more likely to have a higher Social Status. Social Balance theory fits well in friendship and trust networks, but it is not so evident in communication networks. They also found that people connected to a person C are more likely to have the same type of relation with C if he spans a

structural hole, and especially if they are not connected among them.

Searching for evidence to some signed network theories, Lescovek et al. [57] observed that Balanced Theories are actually present in mutual relation networks (like friendship networks), but their evidence is not found in directed relation networks (like hierarchical ones), whose dynamics is ruled by Status Theory. They found strong evidence of the Triadic Closure property and they noticed that positive ties are more likely to be clumped together, while negative ties tend to act like bridges between positive islands (what they have called *embeddedness of positive and negative ties*). From this last evidence, they stated that "all positive triads" are much more common than the "two negatives and one positive triad", from Balanced Theory.

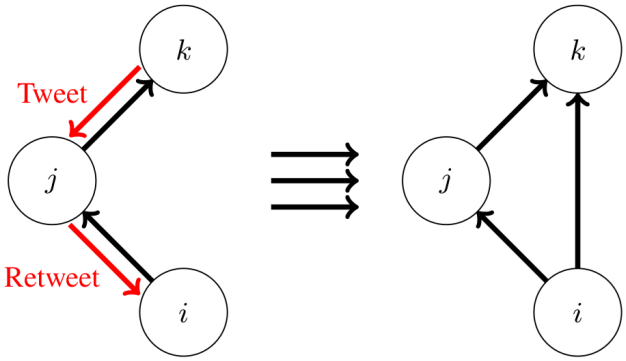


Figure 2.11: Example of Triadic Closure in Twitter [69]. User *i* follows *j*, who follows *k*. When *k* posts a tweet and *j* retweets it, *i* gets exposed to the tweet and there is an increased likelihood of *i* to follow *k*.

About the way dynamics changes the Structural Network, Myers and Lescovek [69] observed that Twitter is highly dynamic in what concerns follow relations and about 9% of all connections change in a month. The rate of new follows and unfollows strictly increases with the user's degree. An average user with 100 followers gains 10% new followers and looses 3%, per month. For Networks built from retweets or mention relations, their dynamics are also topic-dependent, according to Cha et al. [16]. They also found that retweet/mention relation graphs reveal more their Influence properties than the follow relation graphs.

Some of these studies are based on large-scale data that contain millions of users and billions of relations. Representing and analyzing graphs with this magnitude can be computationally demanding. Watanabe and Suzumura [105] analyzed the average degree of separation among Twitter users in a network of 469.9 million users and 28.7 billion relationships. They used the WebGraph Framework [3] to generate and process the inherent network graph of their data. WebGraph is a Framework that exploits networks' redundancy to compress its representation in memory, and includes a set of algorithms and tools to manipulate and analyze large-scale graphs [14]. This framework provides algorithms for clustering, such as PageRank and Layered Label Propagation [13].

2.2.3 Event Detection

Topic trends are topics which are generating a huge flow of information during a certain period of time, for instance a terrorist attack, the beginning of an earthquake, the evolution of political elections [1], and they can contain useful information.

Due to the noisy information that is present in Twitter it is important to filter it. Alonso et al. [1] tried to find some useful indicators to identify tweet's possible interest without analyzing its content, in order to

[3] `http://webgraph.di.unimi.it/`

improve twitter filtering. They exploited *crowdsourcing* in order to manually classify a dataset of tweets and then identify common properties among the interesting ones. They found that tweet content length and the presence of URL can be good indicators of the interest of a tweet. However, they realized that even to humans a large sample of tweets is needed in order to understand tweet relevance, because it can depend on the current context. *Crowdsourcing* labeling is also a good technique to build training sets of labeling data for automatic tools. The downside of their experience is that manual labeling is the subjective nature of the process. Furthermore, the fact that people need large samples to understand the context can also increase the boredom of the task. For instance, an existing system that tracks trending topics about sports in real-time is National Basketball Association – professional basketball league in North America (NBA) Pulse [4]. They use global trending on Twitter to trace big moments, build rankings and statistics about what is happening in the basketball world, with special attention to NBA.

A very interesting example of Event Detection is presented by Sakaki et al. [76] to detect the occurrence of earthquakes using Twitter. They identify three essential characteristics for event detection: (i) it is a large scale event that is experienced by many users; (ii) they particularly influence people's daily life; (iii) and they have both limited spatial and temporal regions, which is important for real-time detection. They assume that each user is a sensor (and there are millions worldwide), who can detect an earthquake and alert it with a timestamp and sometimes also with geo-located information. They realized that very little information diffusion takes place on Twitter in such scenario and the "sensors" are independent and identically distributed. They are continuously performing sentiment analysis to real-time tweets to detect when users are experiencing an earthquake. This analysis can be improved by classifying the tweets' positivity and negativity. They hope to improve their filtering mechanisms as well as their spatial models. Nevertheless, their results are promising, having detected 96% of all earthquakes where 80% were promptly detected for an intensity scale of 3 or more. However, both values scaled up to 100% when the intensity was 4 or more, during their period of tests. Yin et al. [112] did a similar research in Australia and detected earthquakes with one minute delay.

2.2.4 Event Prediction

Event Prediction is an extension of Event Detection since it uses it in order to learn and train the prediction model. This technique has a high economic potential, especially in stock market and sales prediction. They combine Natural Language Processing for Sentiment Analysis, Machine Learning and Statistics to identify patterns and detect them [5]. It shares similar problems with Event Detection, regarding the lack of labeled data to train Sentiment Analysis methods.

Some studies have tried to build prediction models for electoral results, but the results have been contradictory. Jungherr [52] found no satisfactory relation between the popularity of parties and the number of votes, in the 2009 Federal Election in Germany. As example, he claims that Twitter worked as a voice for opposition because the number of retweets of tweets made by opposition parties was superior to the ones regarding the winning party. Contradictorily, another study also about the 2009 Federal Election in Germany conducted by Tumasjan et al. [92] considered that activity in Twitter reflected the elections result. In this case, the approach followed was not based in retweets, but focused in conversations. They observed that Twitter was not only used to communicate information but also to discuss it. They claim that it was possible to find similarities in communities of similar parties. So, they consider Twitter activity a valid indicator of political opinion.

Zhang et al. [113] tried to find a correlation between social media and *e-commerce*. Their observations show that 5% of the eBay [5] query streams have strong positive correlation with topics in Twitter

[4] `http://www.nba.com/pulse/`
[5] `http://www.ebay.com/`

and in the case of trending topics, it rises to 25%. They found that some categories are more likely to generate a stronger correlation, like video games. They found that sport events also have a good correlation, but with some lag between the peak of the trend in social media and the peak in shopping queries (characteristic that could be useful for predictive models).

For predicting movies' success, Asur and Huberman [7] used subjectivity and the positivity/negativity ratio about the movies. $Subjectivity\ ratio = \frac{\#tweets_+ + \#tweets_-}{\#tweets_{neutral}}$, $Pos/Neg\ ratio = \frac{\#tweets_+}{\#tweets_-}$. Sentiment information revealed to be much more valuable after the movie release than before because the subjectivity tends to increase after the release.

Cheng et al. [18] went further and tried to build a model to predict the life of an information cascade. Their initial hypothesis was that large cascades are rare and that the eventual scope of a cascade may be an inherently unpredictable property. They analyzed chains of photo sharing in Facebook and their evolution over time, using a variety of Machine Learning methods, including linear regression, naive Bayes, Support Vector Machines (SVMs), decision trees and random forests. Observing how far cascades reached and their distribution, they found that the cascade size distribution follows a power-law with exponent $c \approx 2$. They identified four major factors that shape cascade growth: content, original poster characteristics, structure of the network and temporal features. They found that the importance of features depends on properties of the original upload as well: the topics present in the caption, the language of the root node, as well as the content of the photo. Correlation between the number of shares and the rate of views of the uploaded photos is higher for those with a Portuguese-speaking root node as opposed to an English one. They also state that larger cascades are more difficult to predict.

2.2.5 Information Flow

Understanding how information propagates among social networks has a great interest for example in Public Health. Today, with air travel, epidemic diseases can spread very quickly. It is critical to understand what is the dynamics of information dissemination during important global events in order to find a better way to spread useful information about the diseases to reach everybody in the general public in case of a pandemic disease. Kostkova et al. [53] realized that during the swine flu 2009 pandemic, the most reputable and trusted media, such as the British Broadcasting Corporation (BBC) [6], played the central role in Twitter information propagation. One limitation of using Twitter is that it is mostly used by younger generations (between 18-29 years old in 2010 [53]).

Tagging is a useful feature of Twitter that helps to disseminate information in target circles. When people are using hashtags, they are directing content [44]. Their tagged tweets will reach everybody that is talking about the same tag, as well as people who are only searching for it.

Information propagation not only depends on the network structure and content, but also on some other characteristics, such as language. Eleta [25] noticed that during the Arab Spring or Awakening the English language acted like a bridge, allowing it reach out to the whole world.

Honey and Herring [42] divide Twitter users into three categories: *info sources*, who post news; *friends*; and *info seekers* who usually do not post anything but follow others' posts. They studied conversations and collaboration via Twitter and found that the average number of participants in a Twitter conversation was 2 in a distribution from 2 to 10 participants. The average length was 25 minutes and 33 seconds in which the shortest conversation was 25 seconds long and the longest 54 minutes and 22 seconds. Distributed from 2 to 30, the average number of messages exchanged was 3 and the average time between messages was 4 minutes and 24 seconds, considering an existing range between 25 seconds and 34 minutes and 5 seconds.

[6] `http://www.bbc.co.uk/`

2.2.6 Influence and Homophily

Influent people have an important role in their network. Finding who are the most influential actors in a given network has a great interest for several areas such as marketing, politics, entertainment.

Having an influential position in the network is hard achieve but it has such a high economic value that some people pay to rise their number of followers. Stringhini et al. [78] detected the action pattern of this type of users and their strategies to create a false image of popularity. Two common methods are: offering some benefit to real users if they follow them (which is frequently used by known brands in promotional giveaways); and creating several fake accounts to follow back (which is more time consuming and less effective).

However, Anger et al. [3], Weng et al. [108], and Cha et al. [16] claim that the number of followers of a user reveals little about his influence in the network. Instead, Anger et al. [3] focus on the number of retweets and mentions, which are more content-oriented and suggests more influence. They call the influential users, alpha users, and they calculate a user *Social network potential* $= \frac{retweetratio+replyratio}{2}$. They found that the top 10 alpha users in Austria are journalists and media. They identified other tools to measure influence, but due to its economic value, their algorithms are not revealed. They aim to improve their work with sentiment and temporal evolution analysis.

The Web is a virtual environment where customers are able to discover a lot of information about products before buying them [41], by word-of-mouth. This plays an important role today in brand's communication [49]. In order to exploit word-of-mouth effect Bakshy et al. [9] measured user influence analyzing some URL shares and checking who-follows-whom among the users who have shared it. Then, considering tweet timestamps, they found who was in the beginning of the cascade and everyone's influence was measured in relation to their distance to the beginning of the cascade. With this, it was possible to predict user influence for some future action, based on his profile and past actions on cascades (and also, considering the current context). Nevertheless, word-of-mouth marketing in Twitter is still hard to implement, because users do not reveal much information about themselves. Existing solutions to infer gender, age, occupation, and interests of users require manual labeling, which is hard and time consuming. Their solution followed two steps: first they checked if the user had some blog URL in their profile: if he had some, they extracted possible additional information from his blog; then they inferred several characteristics from user's neighbors, regarding Social Influence and Homophily theories.

These studies are mostly focused on finding influential actors, due to their direct economical value in viral propagation of information and behavior. However, Influence is believed to exist not only caused by a few number of highly influential nodes, but also by the prevalence characteristics in the network itself. Influence and Homophily shape the network's dynamics [80] and this can be valuable to explore these phenomenons alongside with other types of analysis. Weng et al. [109] report the existence of Homophily in Twitter communities. Ye and his colleagues [111] built a Social Influenced Selection Model in order to capture social influence between linked friends and user preferences, exploiting Homophily to improve recommendation systems. Tang et al. [80] analyzed different community detection methods, and they state that Homophily can be used for community overlapping detection.

2.2.7 Sentiment Analysis: Positivity, Negativity, Neutrality

Understanding the inherent sentiment of each interaction in the network is essential for understanding its dynamics. Sentiment shapes the evolution of the network's structure, and sometimes may be used to reveal the true meaning of what is said. For that reason, several research works in Social Networks include Sentiment Analysis.

In Twitter, this is a particularly difficult task because each tweet does not have its sentiment explicitly stated. Actually, it has to be inferred by the text, smiles, from the user behavior, and from the surrounding context. The problem is that available data is short, biased, noisy, possibly false, and it is particularly difficult to obtain the sentiment of a short message, with few contextual information and lack of training data, using existing sentiment analysis solutions (mostly based on Natural Language Processing (NLP)). Some other variables can make this even more difficult, such as sarcasm.

Several strategies could be combined to get the subjectivity/polarity of tweets or relationships, and there are two categories of classifiers: Supervised and Unsupervised Classifiers. Supervised methods use training models and Machine Learning techniques, depending on training data, while Unsupervised methods rely on static classification of sentimental words and lexical evaluation techniques.

For instance, Li et al. [58] and Hodeghatta [41] chose a supervised approach and built Maximum Entropy classifiers, based on trained prediction models. Joshi et al. [51] also built a supervised sentiment analyzer for micro-blogs, which uses the lexical resource SentiWordNet [7] [26] [8] to train the classifier and detect the polarity of the messages, after a sequence of text normalization steps. Chowdhury et al. [20] also used SentiWordNet as training set for sentiment labeling.

A graph-based solution is described by Jiang et al. [50], in which after tweet normalization, they classify the tweet considering also other related tweets, enriching the classification with a possible context. Also following a graph-based approach, Wang et al. [103] centered their analysis on hashtag sentiment specifically. The reason for using hashtags is that they are strongly associated with topics, from which it is possible to extract the context and detect the sentiment associated. The results have shown to be better when there is a co-occurrence of hashtags because their relationship can enrich the sentiment analysis. The problem is to extract the literal meaning from the hashtag. The first step is to classify the tweets' subjectivity, filtering out the neutral ones, and then assigning the polarity of the remaining ones. This is done with a two-staged (subjectivity and polarity) SVM, which they claim to be a robust strategy for dealing with biased and noisy data. They found that combined hashtags have more likelihood of sharing the same polarity. Exploring this, a relation graph of hashtags co-occurrence is built, in which their sentiments can be related and inferred.

Asiaee et al. [6] also separated neutral tweets (informative ones, like news) from the dataset and pre-processed the sentiment classification normalizing the text, removing retweet and username references and URLs. Then they represented the tweets according to the Bag-of-words model and pruned the highly frequent words. The training set was labeled manually for different topics using *crowdsourcing*. A sparse modeling approach was used to do the classification and three supervised techniques were tested: SVM, K-nearest neighbors and Naive Bayes. From these classification methods, the SVM produced better results.

Most of the methods are content-centric, but the approach used by Hu et al. [43], tries to go beyond content, applying Information Flow and Contagion Theory to sentiments. So, they consider emotional contagion and propagation of happiness to improve sentiment classification. They used the Stanford Twitter sentiment dataset [32, 33] from Sentiment140 API [8], even though, they state that there is a wide lack of manually labeled training data for sentiment analysis in microblogging environments, such as Twitter.

The major problem of supervised techniques for Sentiment Analysis is the need of manually labeled training data, which must be adapted to the social environment and the topic in analysis. Lexicon-based techniques do not need to be trained and their result is context independent.

Two applications of such unsupervised methods are the tools SO-CAL [79] and SentiStrength [9].

[7] http://sentiwordnet.isti.cnr.it/
[8] http://help.sentiment140.com/api
[9] http://sentistrength.wlv.ac.uk/

Twitter	Positive Correct	Negative Correct
Unsupervised SentiStrength	59.2%	66.1%
Supervised SentiStrength	63.7%	67.8%
Best machine learning	**70.7%**	**75.4%**

Table 2.2: SentiStrength evaluation results for Twitter data [89]. Metric used: accuracy regarding the golden standard created by 3 human coders. Comparison between Unsupervised and Supervised SentiStrength and the best result of different machine learning techniques used.

Weber et al. [107] used SentiStrength to classify the sentiment of tips' comments in Yahoo Answers [10]. Durahim et al. [23] analyzed Turkey's overall sentiment during the first quarter of 2014, using SentiStrength over a dataset of 35 million tweets collected from Twitter. SentiStrength uses a lexical approach for sentiment classification and it is designed for social web texts [89]. This tool is available for purchase for commercial users like Yahoo, and free for researchers and educational users. The SentiStrengh's algorithm outputs two independent values: a positive and a negative score from 1 to 5 and -1 to -5, respectively. For each text, words, emoticons and punctuation are splitted. Then, for each word it searches for a match in a lexicon dictionary of sentimental terms, labeled with a positive or negative value. The overall score of a sentence is the maximum value of all positive words in the sentence and the minimum value of all negative words, with the default result of 1 and -1. Beyond a sentiment term strength lexicon, there is also a dictionary of emoticons and another of idioms. The algorithm includes some other rules such as bipolar values of sentiment for special cases of words with dual polarity, spell correction and strength boost for repeated letters, list of booster words to strengthen immediately following sentiment words, negating word list to neutralize following sentiment words, exclamation marks and repeated punctuation strength boost, capital letters strength boost, strength boost of both consecutive moderate/strong positive or negative terms, and a list of typical ironic sentences [83]. Although the original lexicon was built for English, there are already adapted versions for other languages such as Spanish, Portuguese, German. The dictionaries and lists files are modifiable and rules can be adapted or disabled, which can be useful, not only for language modification, but also for specific topic adaptation. SentiStrength also includes a supervised mode that updates the original lexicon and rules according to the provided training set [83]. The authors evaluated the accuracy of SentiStrength [89] using different datasets, including a Twitter dataset as is shown in Table 2.2. The accuracy was measured for unsupervised and supervised modes and compared with the best result of all machine learning algorithms tested: SVM (Sequential Minimal Optimization variant, SMO), Logistic Regression (SLOG for short), ADA Boost, SVM Regression, Decision Table, Naïve Bayes, J48 classification tree, and JRip rule-based classifier. The golden standard considered was result of the evaluation of three manual coders. The major discrepancy was found between the unsupervised method against best machine learning approach with SentiStrength scoring 11.5% less for positive sentiment classification. The accuracy increases for negative values and there is a slightly increase using SentiStrength in supervised mode.

Despite the quality of the used technique, Li et al. [58] stress out that another open problem in Sentiment Analysis is the lack of studies that relate the tweet sentiment and the real humor of the user.

2.2.8 Spam Filtering

Irrelevant content (spam) in social contexts is mostly caused by automatic applications, commonly known as bots, which are frequently used for intensive commercial advertising propagation. This type of content has few interest because it does not represent real social interactions and it is usually out of context.

[10] `https://answers.yahoo.com/`

Spam decreases the quality and the veracity of the datasets analyzed and it is pointed out as a possible source of problems that can bias the results in the majority of the studies in Social Network Analysis. Although several techniques are used to minimize the amount of spam inside the datasets, even to humans it is hard to properly identify whether a tweet is spam [1] or not.

One simple but also subjective and time consuming solution suggested by Kostkova [53] is finding the most frequent shared URLs in the dataset and manually classify them. Then, all tweets that include URLs labeled as spam are removed.

Another user-focused solution given by Chu et al. [21], focus its attention to profile properties. They classify users as *cyborg*, *human* and *bot*, in which cyborg is an occasional user who allows applications to do automatic tweets during his inactive period. First, they check if the users have their profile protected and if they are verified accounts (like Bill Gates' account, which certificates its authenticity). Then, they use Probability Models such as High Entropy Models and Bayes Theorem to classify the users according to their account properties. They improve spam detection by checking shared URLs in Google Safe Browsing blacklist [11]. According to their results, they observed that user accounts classified as bots share an URL in 97% of their tweets. Those accounts classified as human show an average ratio of 29% on sharing URLs in their tweets.

2.2.9 Geo-location

Many researchers point that location information can enrich their solutions in future work. However, this is a problem by itself due to the amount of false data provided by the users in their profiles' field of location [53]. The lack of accuracy currently observed in the location information is stated as the main reason for not being widely considered yet. However, there are some strategies that can help to infer user location, beyond their profile information:

Li et al. [58] believe that an aggregate analysis could help to infer users' true location. Following this same approach Hecht et al. [40] found that 34% of users did not provide real location information on their Twitter profiles, and when they did almost never detailed more than their city. They performed simple Machine Learning experiments that found trustful information about certain users (with coordinates or apparently correct data) and used them to train their models. Then they predicted users' location based on their behavior.

Lin [60] collected 180 million geo-coded tweets about Boston Bombings during one month. The goal was to understand the geo-distribution of fear and solidarity towards this disaster. Social support was easily identified with the hashtag "#prayforboston", which is apparently a variant of other similar hashtags "#prayfor$\{...\}$" used during other disasters worldwide. It was found that communities that expressed more fear also expressed more solidarity with Boston. Geo-social close cities dominated the top of solidarity group, with prevalence from the US. On the opposite end, London, for example did not reveal strong solidarity.

According to Cuevas et al. [22] most of Twitter users perform their activity in an area of at most few kilometers covering few cities, within an unique country. They define locality as the phenomenon that makes the activity and/or relationships of a user in Twitter to remain local within its geo-cultural-political community. It was found that language affects locality. Brazil was the country showing highest locality (80% of local links) since it is a big country with a strong culture and a single spoken language, which has the biggest number of speakers of that language on Twitter. And also because other countries with the same language are not representative in Twitter, which is the case of Portugal. Spain, on other hand, has only 41% of local links.

[11] `https://developers.google.com/safe-browsing/`

2.2.10 2010 FIFA World Cup on Twitter

Football has millions of fans worldwide and football competitions generate big bursts of online conversation and discuss [29] [97] [71]. These competitions are sources of large amounts of data and this data is characterized by sharing a common general topic and belonging to a same specific period in time.

Nichols [71] collected a dataset of tweets from 36 games of the 2010 World Cup. Most games were recorded using the keywords "worldcup" and "wc2010", promoted by FIFA, through Twitter's Streaming API [100]. Sporting events consist of a sequence of moments, each of which may contain actions by the players, the referee, and the fans. They use collective information from several users to identify important moments within an event and also the event itself. This event detection is based on two properties of the Twitter stream: sudden increases, or "spikes", in the volume of tweets in the stream suggest that something important just happened because many people found the need to comment on it; a high percentage of the contents of the tweets at a "spike" in volume contain text describing what happened in the moment, and this text often contains repetitive elements, such as the names of players involved, the type of the event.

Important moments can often be detected in Twitter streams, when the volume of status updates increases sharply. Over the course of a sporting event, this may happen many times. They discovered two problems in using the absolute value of the volume to detect important moments: sometimes the stream volume may stay high for several minutes and have localized peaks; and some moments generate significantly less traffic than others and they might be missed. To avoid these problems, they have chosen to use a detection algorithm based on the change in volume (i.e., the slope of the volume graph). This algorithm is based on spike detection which, having tweets from the entire event available, computes a threshold for the entire event using basic statistics about the set of all slopes in that event.

For example, a threshold for a particular soccer match may be computed from tweets recorded for that entire match. The first threshold tried was $median + 2 * standarddeviation$, but it could not detect smaller spikes. Then they adopted a threshold of $3{*}median$, which produced results that closely matched a visual inspection of spikes across their 36 game dataset. After identifying all slopes that exceeded the threshold, they generated a list of "spikes" that correspond to the important moments in the event. Each spike can be defined as a tuple of $<$*Start time, Peak time, End Time*$>$.

For each slope above threshold, they calculated the start time by searching backwards in time until they found the point where the slope began going up. The peak time was calculated by searching forward until the point where the slope started decreasing. The end time was calculated by searching forward from the peak until they found the point where the slope begins increasing again. Before returning the list of spikes, any duplicates were removed, which may happen for large spikes that include multiple above threshold slopes.

They used these techniques to build a system that produced automatic summaries of sporting events according to the detected spikes of content. These summaries are generated using the most frequent sentences during the spike and they usually describe moments such as goals, faults, injuries, penalties.

2.2.11 Twitter as a mirror for other Social Environments

Even though these examples are strongly related with Twitter, most of them are probably repeatable in other social environments. Shuai et al. [77] compared the detection of popular events in Twitter and Weibo [12], which is a Chinese microblogging website similar to Twitter. Their purpose was to identify differences in response to commonly interesting popular events, such as the 2012 US presidential election, considering their different backgrounds and different traditions, since Twitter is blocked in China. The re-

[12] `http://www.weibo.com/`

sults have shown that they both share a similar degree of interest towards commonly interesting events. The response to hot events had at most one day of delay. They both also shared a similar degree of popularity, temporal dynamic and information diffusion patterns. The major differences identified were related to the origin of the information propagation. While Twitter networks were "infected" internally, starting from content produced inside the network, Weibo dissemination began with external sources, such as news, or other websites.

2.3 Combining Community Detection, Sentiment Analysis, Influence and Homophily

So far we have seen different studies with relevant outcome for our work, especially on data extraction, graph representation, clustering, sentiment analysis and some interesting findings in influence and homophily analisys. However, we did not found any study that combined clustering techniques and sentiment analysis for finding influence and homophily evidence. The three following cases sustain the motivation of our work, finding in the work of Fowler and Christakis [31] evidence of sentiment homophily based on survey data, while Thelwall [84] explored sentiment homophily between pairs of friends using social media data, and Gruzd et al. [38] extracted a topic-oriented dataset related with a big sporting event and used it to find clusters and perform sentiment analysis.

Fowler and Christakis [31] conducted a study about the propagation of happiness within social networks, using data from the Framingham Heart Study [13] collected between 1983 and 2003. From this data, they extracted a network of $5,124$ individuals and $53,228$ respective social ties. Each person was weekly asked how often they experienced certain feelings during the previous week: "I felt hopeful about the future", "I was happy", "I enjoyed life", "I felt that I was just as good as other people". They used this information to measure the state of happiness of individuals throughout a period of time. According to their results there is happiness homophily in clusters with up to three degrees of separation between nodes. They also had information about people's address, which allowed them to find that geographic proximity among connected people increases the probability of sharing the same state of happiness. This study not only found evidence of sentiment propagation through influence, it also suggests that it may cause sentiment homophily at a cluster level.

Thelwall [84] searched for homophily in social network sites using data extracted from MySpace [14], concluding that there was a high significant evidence of homophily for several characteristics such as ethnicity, age, religion, sexual orientation, country, and marital status. Then, based on the same type of data he conducted another study on [85] emotion homophily. Using an initial version of SentiStrength for sentiment classification, two different methods were tested to seek emotion homphily between pairs of friends: direct method and indirect method. The direct method compares only the sentiment of the conversational comments between each pair of friends. The indirect method compares the average emotion classification of comments directed to each node in each pair of friends. Weak but statistically significant levels of homophily were found for both methods. However, the direct method can only give insight of the average homophily at a maximum distance of 1, while the indirect method covers a maximum distance of 3. Both methods do not take into account cluster configurations in the network and the covered range of time considered in the analysis is not specified.

Gruzd et al. [38] followed the study of Fowler and Christakis with web-based social network data, focusing on the potential propagation factors for sentiment contagious instead of searching for evidence of sentiment homophily. They performed a topic-oriented data extraction from Twitter in order to minimize

[13] Medical study about cardiovascular disease – `https://www.framinghamheartstudy.org/`
[14] `https://myspace.com/`

possible bias caused by the occurrence of multiple events that generate multiple unrelated discussions, and they found on the 2010 Winter Olympics a well covered and very popular event on Twitter, from which they got strong emotional content. Using SentiStrength for tweets' sentiment classification, they found that a tweet is more likely to be retweeted through a network of follow relations if its tone and content are both positive. Fan et al. [30] decomposed sentiment into four emotions: angry, joyful, sad and disgusting. They used a bayesian classifier to infer these emotions based on emoticon occurrence in interactions extracted from Weibo. Considering pairs of direct friends in a follow-relation network, they only found evidence of emotion homophily regarding anger and joy, observing that anger was the most influential emotion and the chance of contagion was higher in stronger ties. Using a follow-relation network extracted from Twitter, Bollen et al. [15] also found sentiment homophily but regarding sentiment polarity, which they called subjective well-being assortativity. They observed that pairs of friends connected by strong ties are more assortative, however they did not identify whether this phenomenon was caused by selection or social influence. None of these studies analyzed sentiment dynamics over time nor looked into an overall sentiment at community level.

2.4 Summary

With the several studies discussed in this Chapter we understood that Twitter offers good solutions for data extraction; there were found different types of relations in Twitter; there are solid techniques to represent, manage and cluster social networks' graphs from those relations; it is also possible to analyze the inherent sentiment of tweets even though without a strong accuracy rate; homophily and influence were found in social media relations for certain features; and there are few studies that relate sentiment with influence and homophily but they found evidence of a possible correlation. This way we realized that there are traces of sentiment homophily, however this phenomenon is not studied at a cluster level. Following this open question, we propose to look for signs of sentiment homophily in social circles and understand whether prevalent sentiment in communities can be used for estimating individuals' sentiment.

According to the topics covered by those studies our work must comprise:

- Twitter data extraction and analysis;
- Network representation and clustering;
- Sentiment Analysis;
- Influence and Homophily analysis.

We found different approaches and solutions for each topic, depending on the author and the purpose of his work. In Table 2.3 we summarize those different decisions.

For Twitter data extraction most authors used the Streaming API [71, 105, 83, 55] which allowed to extract a collection of tweets continuously during a period of several months. To build retweet/reply relation graphs this API is able to provide a high volume of data. Gruzd et al. [38] decided to build a follow relation graph and to do that he used the REST API. The major disadvantage of this API is having strict limits in data retrieving, providing smaller datasets. Nichols et al. [71] and Gruzd et al. [38] stated that performing a topic-related data extraction of big sporting events such as the FIFA World Cup and Winter Olympics generates datasets with high volume of data and reduced amount of noise. Tang et al., Cha et al. and Huberman et al. [80, 16, 45] use retweets and mentions to represent the networks' graphs because they consider that these relations represent strong ties while follows are weak ties.

With large amounts of data it is possible to infer large networks. To deal and manage these types of graphs Watanabe and Suzumura [105] used the WebGraph framework. This framework created by Boldi et al. [14, 13] is able to be used with another framework that provides an implementation of the Layered Label Propagation algorithm that can be used for clustering.

Thelwall et al. [88], Lai [55], and Gruzd et al. [38] used SentiStrength to perform tweets' sentiment classification showing good results in unsupervised mode, which becomes a suitable solution to classify the sentiment of texts for topics with lack of training data, instead of using supervised techniques.

Thelwall [85] found sentiment homophily among direct pairs of friends, while Fowler and Christakis [31] state that there is sentiment homophily in up to three degrees of separation between friends. However, the dataset characteristics between these two studies are quite different. The first uses social media relations and analyzes their inherent sentiment with an automatic classification tool, which gives instantaneous insights about individuals' mood which are particularly ephemeral in time. The second uses surveys where individuals classified their overall mood over the previous week. This way, with social media we are able to analyze people's sentiment in narrower time windows, however it does not give information about their own perception about their overall sentiment during larger periods of time.

Our approach includes some decisions, tools and methodologies that have revealed good performance and lead the several cited studies to reliable results. Namely, we chose to extract our data from Twitter using Twitter Streaming API and sampling a dataset about the sporting event 2014 FIFA World Cup; build the network graph according to retweet and mention relations, which mirror strong ties on Twitter; for the representation of the graph we chose to use WebGraph framework which includes the Layered Label Propagation algorithm that is used for clustering; SentiStrength is widely used in the literature, offering a versatile solution for sentiment classification with no demand of training data and good accuracy results for Twitter data.

Author	*Twitter data extraction*	*Network relations*	*Sentiment Analysis approach*	*Influence and Homophily Analysis findings*
Nichols et al. [71]	**Twitter Streaming API. Topic-based extraction. High volume of data for big sporting events: FIFA World Cup.**			
Watanabe and Suzumura [105]	**Twitter Streaming API. 3 months data extraction.**	**WebGraph framework for large-scale dataset.**		
Boldi et al. [14, 13]		**WebGraph framework for large-scale dataset. Layered Label Propagation algorithm for clustering of large-scale graphs.**		
Tang et al. [80]	**Following: weak ties, Retweet/Mention: strong ties.**			
Cha et al. [16]	**Following: weak ties, Retweet/Mention: strong ties.**	**Retweet/Mention relation graphs.**		
Huberman et al. [45]	**Following: weak ties, Retweet/Mention: strong ties.**	**Retweet/Mention relation graphs.**		
Thelwall et al. [83, 87, 85]	**Twitter Streaming API.**	Follow relation graph.	**Lexicon-based classification. SentiStrength for Twitter.**	**Sentiment Homophily between pairs of friends.**
Lai [55]	**Twitter Streaming API.** 6 months data extraction.		**SentiStrength for Twitter. Sentiment-based topic detection. Time-series sentiment analysis.**	
Fowler and Christakis [31]		Real-life friendship, family, spousal, neighbor, and coworker relationship graph.	Weekly questionnaire about mood state.	**Sentiment Homophily up to three degrees of separation between friends.**
Gruzd et al. [38]	Twitter REST API. 21 days data extraction. **Topic-based extraction.** High volume of data for big sporting events: Winter Olympics.	Follow relation graph.	**SentiStrength for Twitter.**	Higher likelihood of retweet for positive tweets.

Table 2.3: Relevant contributions suitable to the scope of this work, with comparison between some different techniques and approaches. In bold are represented the ideas and methodologies that we chose to follow in our research.

Chapter 3

Data Overview

3.1 Twitter Developer APIs

With a dynamic flow of 500 million new tweets per day [95], Twitter is one of the major sources of social data available on the Web and it is widely used for social studies as described in Chapter 2. Furthermore, Twitter provides public access to a portion of its data, offering two different public APIs to obtain it in a structured format, which makes Twitter a suitable source of data for our research. Despite the amount of data that is possible to obtain from Twitter, this data can be biased and noisy. There are many unrelated topics being discussed at the same time, and we chose to collect a topic-oriented dataset to grant at least one common characteristic among the majority of the collected data. Nichols et al. [71] and Gruzd et al. [38] reported that big sporting events generate streams of enthusiastic discussions and high volume of content, therefore we chose to use the 2014 FIFA World Cup as topic of extraction, expecting to catch emotional tweets in this stream.

To find social circles and analyze their behavior over time, a large amount of data needs to be extracted during a period of several weeks. Between the two existing Twitter APIs, we found the Public Streaming API to be the most appropriated channel to collect our dataset, because it gives low latency access to Twitter's global stream of public tweets, allowing to maintain a persistent connection, which is suitable for following specific topics, and data mining [100]. REST API is more useful for singular searches, reading user profile information, or posting tweets, having strict rate limits for periods of 15 minutes.

Connecting to the Twitter Public Streaming API requires keeping a persistent Hypertext Transfer Protocol (HTTP) connection open as it is illustrated in Figure 3.1, from which a stream of tweets is received as JavaScript Object Notation (JSON) encoded data [1]. Each tweet is a JSON object and the rate of upcoming tweets is also limited but in percentage regarding the real-time flow amount of occurring tweets.

This API offers two endpoints:

```
1 HTTP GET statuses/sample
2 HTTP POST statuses/filter
```

The first endpoint provides unfiltered tweets as response to a HTTP GET request, while the second returns public statuses that match one or more filter predicates requested through HTTP POST. These predicates can be user IDs, keywords, locations or message lengths.

In addition to the standard JSON encoded tweets, other kinds of messages may be delivered on the stream, such as tweet deletion notices, blank lines, connection warnings, limit messages announcing

[1] http://json.org/

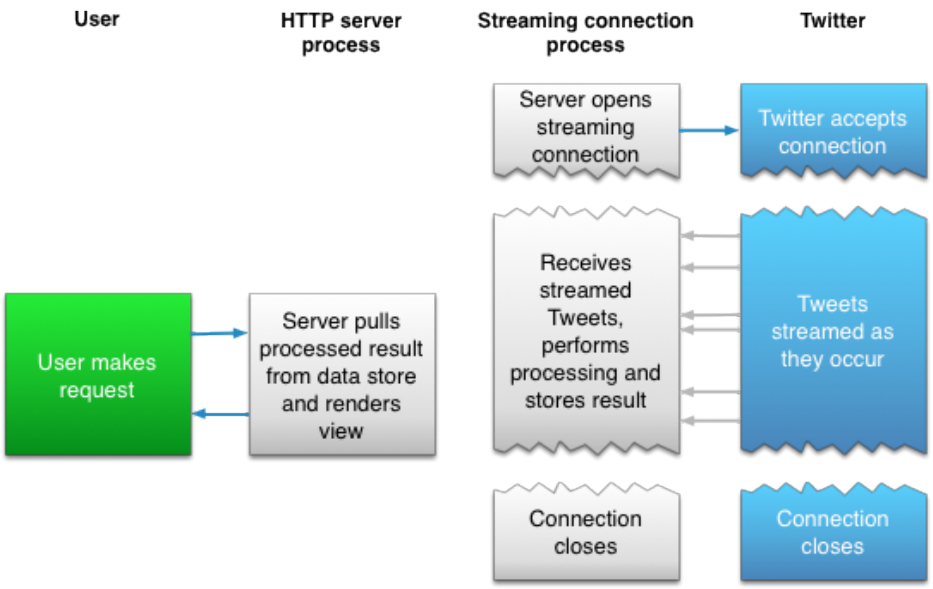


Figure 3.1: Scheme of the HTTP connection to Twitter Public Streaming API [100]

the number of tweets not streamed.

The JSON structure of a tweet depends on the tweet type. Appendix B has an example of each type: simple tweet, retweet and reply.

- A simple tweet does not contains the field **"retweeted_status"** and the field **"in_reply_to_user_id"** is null;
- A retweet contains the original tweet as value of the field **"retweeted_status"**;
- A reply has the receiver ID as value of the field **"in_reply_to_user_id"**.

A tweet cannot be both a retweet and a reply. Mentions, hashtags and URLs can be found in any type of tweet inside the **"entities"** object that is also part of the JSON object. All replies have at least one mention.

3.2 Extracted Dataset

The endpoint used to extract our dataset was **HTTP POST status/filter**, available at `https://stream.twitter.com/1.1/statuses/filter.json` [98]. To handle the HTTP connection we used the Perl module **AnyEvent-Twitter-Stream-0.27** [68], creating an event loop to receive compressed data of streaming tweets. This connection requires OAuth [2] authentication for registered developer users. The received data was filtered using the list of keywords about the 2014 FIFA World Cup, Appendix A, and was stored in zipped files of at most $1,000,000$ messages.

Extraction started on March 13th of 2014 and it ended on July 15th of 2014, covering the entire event that took place from June 12th to July 13th of 2014. It resulted in 166 GB of compressed data, distributed in 419 files, containing a collection of $339,702,345$tweets. Table 3.1 shows the distribution of each type of tweets in the dataset.

We found that 64.7% of all tweets in the dataset have at least one mention, which makes it the most frequent type of strong relation in the dataset, followed by the retweet and finally the reply. However,

[2] `http://oauth.net/2/`

	All Tweets	*Simple Tweets*	*Retweets*	*Replies*
Total	339,702,345	140,613,063	173,966,384	25,122,898
Rate	100%	41.4%	51.2%	7.4%

Table 3.1: Tweet type distribution.

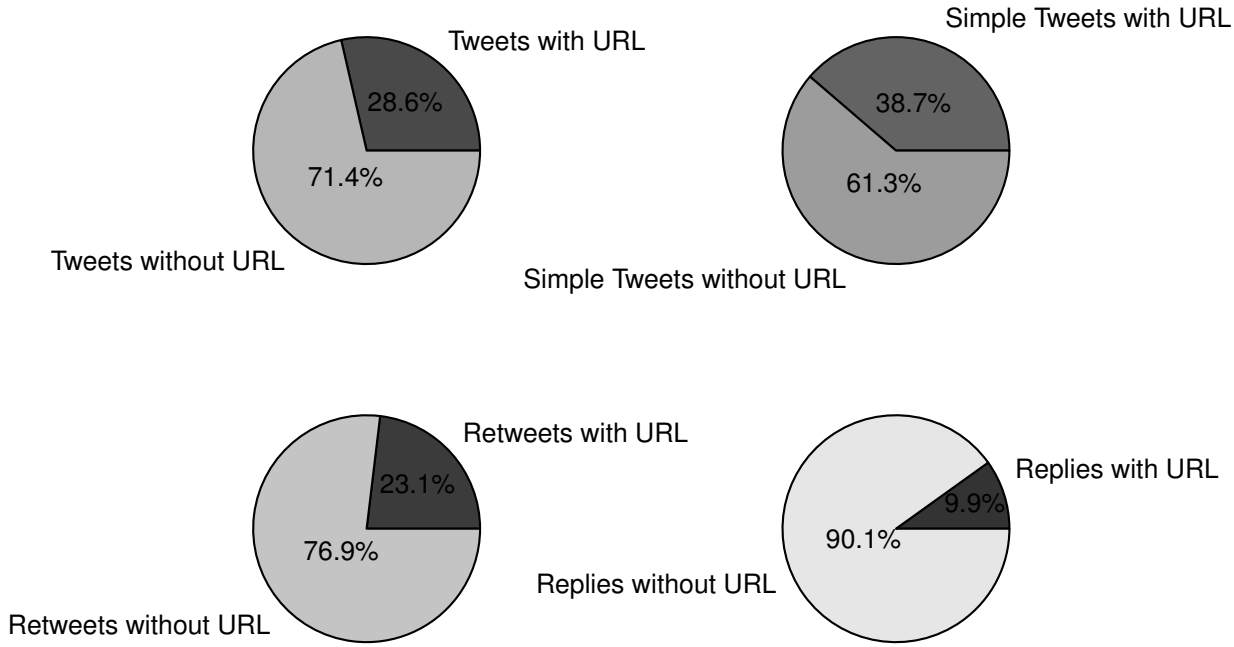


Figure 3.2: Presence of tweets with URL in the dataset.

	All Tweets	*Simple Tweets*	*Retweets*	*Replies*
Total	97,403,564	37,222,855	53,818,351	6,362,358
Rate	100%	38.2%	55.3%	6.5%

Table 3.2: Tweet type distribution in the knock-out stage subset.

the set of mentions contains the set of replies and also intersects with the set of retweets, mixing the three different types of strong ties together. Therefore, For conducting an independent analysis for each type of interaction, it was more valuable to consider only retweets and replies, since they are mutually exclusive.

Regarding the number of tweets sharing URLs, we found that $28.6\%$of the data contains at least one URL. According to the work of Chu et al. [21], humans share URLs in 29% of their tweets, on average, which can indicate that this dataset contains mostly human created content. This percentage is especially lower if we look only for Retweets and Replies as is shown in Figure 3.2.

The field **"lang"** indicates the machine-detected language of the Tweet text, or "und" if no language could be detected. There are 66 different detected languages in our dataset, however 95.4% of all tweets are written in one of the 10 more frequent found languages, as Figure 3.3 shows.

Due to the amount of countries participating in the World Cup, we only considered a subset of the entire dataset for the majority of our analysis. This subset covers the knock-out stage of the event, from June 27th until July 15th, which represents 28.7% of the entire data. We did this to minimize the sparsity of the information, since only 16, from the initial 32 participating countries, were still in competition. Table 3.2 shows the distribution of each type of tweets in this subset, in which is observable a little increase on the average of retweets over simple tweets and replies.

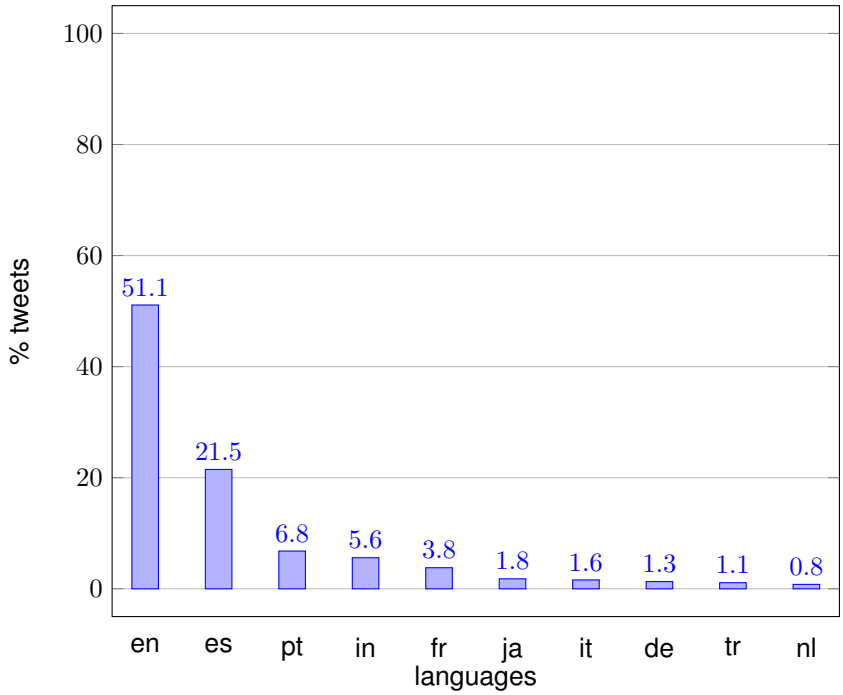


Figure 3.3: Top 10 languages in the dataset: English, Spanish, Portuguese, Indonesian, French, Japanese, Italian, German, Turkish, Dutch.

The knock-out stage comprises the four final rounds of the World Cup, where the following national teams played:

- Round of 16 (June 27th – July 1st): Brazil, Chile, Colombia, Uruguay, Netherlands, Mexico, Costa Rica, Greece, France, Nigeria, Germany, Algeria, Argentina, Switzerland, Belgium and United States.
- Quarter-finals (July 2nd – July 5th): Brazil, Colombia, France, Germany, Netherlands, Costa Rica, Argentina and Belgium.
- Semi-finals (July 6th – July 9th): Brazil, Germany, Netherlands and Argentina.
- Final (including the 3rd place game, July 10th – July 15th): Brazil, Netherlands, Germany and Argentina.

With 6 Spanish-speaking countries in the knock-out stage, from which Argentina played the final against Germany, the winner of the tournament, and with the host Brazil loosing in the semi-final, we observed that the number of tweets in Spanish, Portuguese and German increases 6.3% in this subset, as Figure 3.4 shows. For this reason we chose these three languages to be considered in our Sentiment Analysis alongside with the most prevalent language: English.

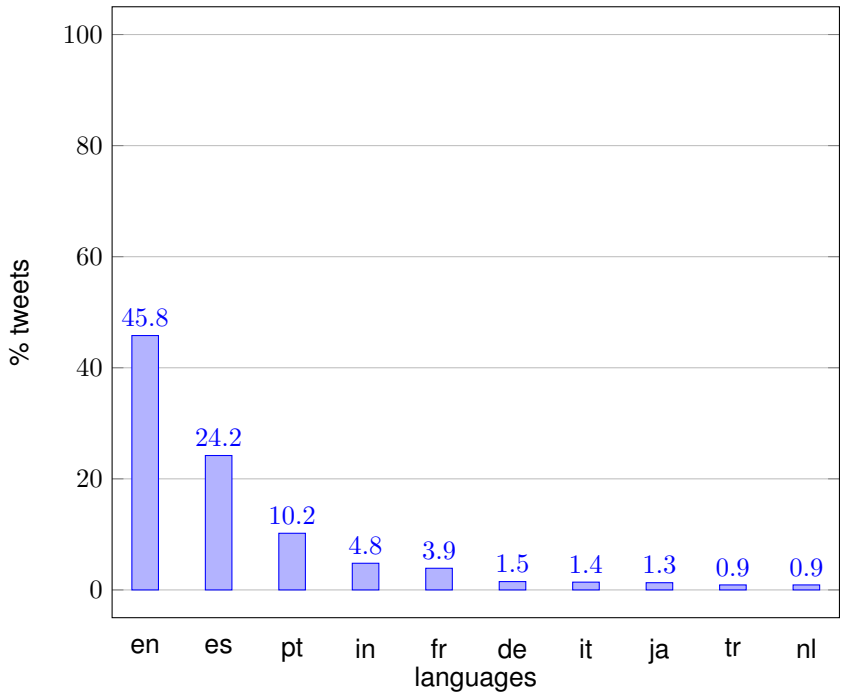


Figure 3.4: Top 10 languages in knock-out stage subset: English, Spanish, Portuguese, Indonesian, French, Japanese, Italian, German, Turkish, Dutch.

Chapter 4

Approach

With this work we aim to understand how Sentiment behaves at a cluster level, how it changes over time and perceive whether it can influence the Sentiment of individuals, inducing Sentiment Homophily in Social Circles. Our approach is divided into four stages: User Clustering; Tweet Clustering; Sentiment Analysis; and Influence and Homophily Analysis in time series, as it outlined in Figure 4.1.

The first three stages integrated existing solutions for clustering and sentiment analysis with several scripts for data transformation. They were used to process the extracted dataset into time-series of sentiment information in social circles. With preprocessed data obtained from these three stages, we defined a set of metrics to evaluate the extent of sentiment homophily. Then, we propose a strategy to ascertain a possible relation between influence and sentiment, which can eventually improve the sentiment classification of tweets in clusters that denote Sentiment Homophily.

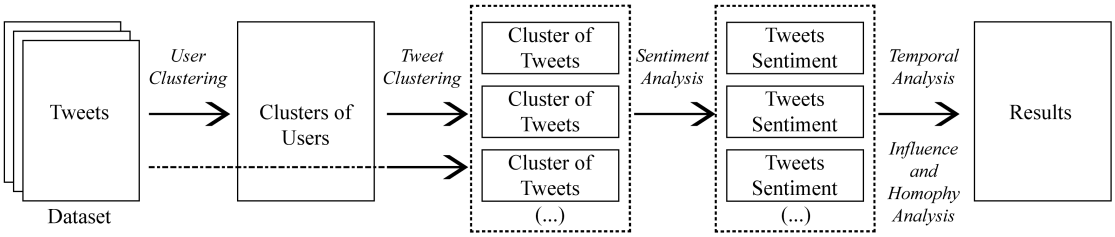


Figure 4.1: High-level view of the designed workflow.

4.1 User Clustering

Before finding the social circles, we needed to find the social network that comprises them. We decided to build the network's graph considering only strong ties, which the literature states to be found in retweets and mentions [80, 16, 45]. However, we chose to use only replies, because retweets and replies are mutually exclusive and replies represent direct conversations, which may not be necessarily true with mentions.

We started by filtering all retweets and replies from the dataset, converting them to a condensed format. Retweets (RTs) have a special field in their JSON structure which is **"retweeted_status"**, that contains the original tweet. Replys (REs) have the ID to whom the tweet is directed in the field

"in_reply_to_user_id". We implemented Algorithm 1 to do this filtering, using Apache Spark framework [1] to parallelize the filtering process and reduce the computation time. Each retweet and reply was transformed from JSON to the format *"type tweetID userID receiverID timestamp"*:

```
1 RT 487733586481938432 187656959 336145436 Fri Jul 11 23:02:03 +0000 2014
2 RE 487733586238242816 97583989 277804136 Fri Jul 11 23:02:03 +0000 2014
```

Algorithm 1 Retweet-Reply Filtering

Require: $Dataset$ of JSON encoded tweets

for each line l in $Dataset$ **do**
  $t \leftarrow json(l)$
  if $retweeted_status$ in t **then**
    print($RT\ t.id_t.user.id\ t.retweeted_status.user.id_t.created_at$)
  end if
  if $t.in_reply_to_user_id$ not $null$ **then**
    print($RE\ t.id\ t.user.id\ t.in_reply_to_user_id\ t.created_at$)
  end if
end for

To analyze the clusters in different periods of time, we filtered and sorted the set of Retweets and Replies by their timestamp values, according to the desired time interval. We also separated Retweets and Replies from each other, for independent analysis.

Since we were dealing with networks with millions of nodes and edges, we chose to use WebGraph [2] to build and analyze their underlying graphs, and used LAW software library [3] for clustering them. These two libraries offer an integrated solution of several algorithms to manage and study large web graphs, over compressed representations. WebGraph works over compressed graphs, for which it offers a compression method to transform the textual format ASCIIGraph to the compact format BVGrpaph.

Assuming n as the number of nodes of the graph, in the ASCIIGraph representation each node is labeled with an integer in $\{0, ..., n-1\}$, and all the nodes to which the node n_i is connected are stored in line $n_i + 2$ with their labels separated by white spaces. The first line of the file has the value of n. We had to transform our set of Retweets and Replies into an ASCIIGraph, so we relabeled the nodes by sorting all the nodes in the list of relations, and assigning them the index of their position in the list. Then, for each relation (n_1, n_2) the new label of n_2 was added to the respective line of n_1 in the ASCIIGraph.

Besides compressing the ASCIIGraph to the WebGraph's format BVGraph, we had to symmetrize it to an undirected and loop-less graph to be used by the LAW implementation of the Layered Label Propagation algorithm, to do user clustering. The symmetric graph was also used to calculate the connected components of the network.

The Layered Label Propagation algorithm [13] is an iterative strategy that reorders the graph such that nodes with the same label are close to one another. This node reordering is useful for graph compression, however for our purposes we only required the node labeling assignment produced by the label propagation algorithm that returns a clustering configuration of the graph. Essentially, each node in the graph is initially assigned with the label corresponding to its own index, then, for each node n, let $\lambda_1, ..., \lambda_k$ be the labels currently appearing on the neighbors of n, k_i be the number of neighbors of n having label λ_i, and v_i be the overall number of nodes in the graph with label λ_i, for a given number of updates, n is updated with the label that maximizes

$$k_i - \gamma(v_i - k_i), \tag{4.1}$$

[1] http://spark.apache.org/
[2] http://webgraph.di.unimi.it/
[3] http://law.di.unimi.it/software.php

in which γ is a function that encodes the resistivity of n to update its label to λ_i. When $\gamma = 0$, the update routine will choose the prevalent label of its neighbors, which originates bigger and sparser clusters, while, for higher γ, the clusters become smaller and denser. The output depends on the initial randomly chosen node where the algorithm starts the labeling propagation. It runs iteratively through a sequence of different γ, which performs a default number of updates.

With both Connected Components and Layered Label Propagation algorithms we get a list of connected component labels and a list of cluster labels, respectively, assigned to the users represented by the index of the list, as in the original ASCIIGraph. The clustering result is mappable with the sorted list of user IDs, and all these steps are outlined in Figure 4.2.

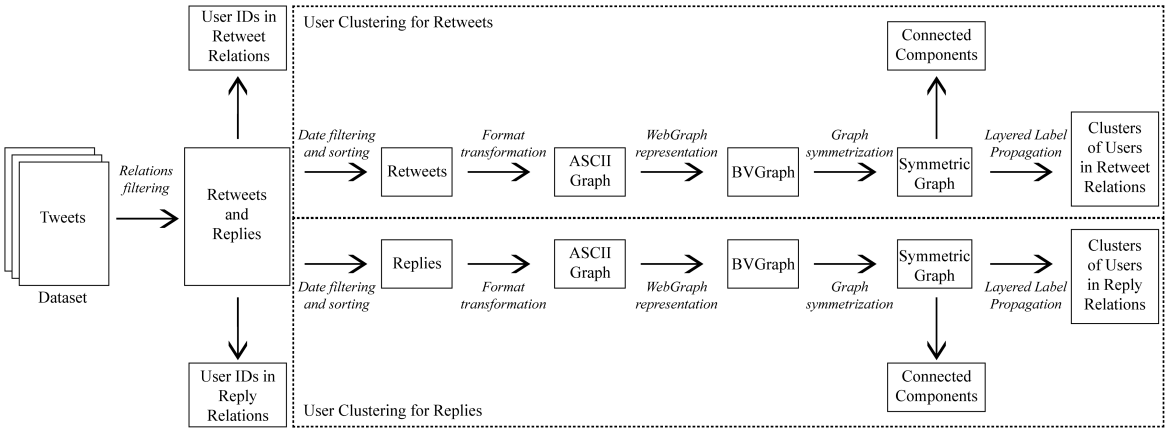


Figure 4.2: User clustering process.

4.2 Tweet Clustering

At the end of the User Clustering stage, we get a list of cluster labels that is mappable with the list of user IDs. With these two lists we are able to know the cluster to which each user belongs to. Our strategy to classify the sentiment of a cluster was getting the tweets that the users in that cluster tweeted during the lifetime of the cluster, and then classify each one, independently, to sum up an overall result. For that, first we extracted from the dataset all the tweets created in the same period of time used to cluster the users, then we converted them to the shorter format *"userID tweetID language epochTimestamp hashtagCounter URLCounter mentionCounter tweetText"*:

```
1 70889888 486688823468769280 en 1404867032 1 0 0 Well now rest of the
    Brazilians can join the anti-FIFA protests. #BrazilvsGermany
```

The obtained subset was sorted by user ID and it was used along with the cluster list and user list to do tweet clustering, according to Algorithm 2.

All the clusters with only one or two tweets were removed. Each cluster of tweets was filtered and divided by its prevalent language, in order to perform sentiment classification without mixed languages. Figure 2 illustrates the process before applying sentiment analysis, where it is shown that only English, Spanish, Portuguese and German were used.

Algorithm 2 Tweet Clustering

Require: $UserIDs$ a list of user IDs, $UserClusters$ a list of cluster labels matching $UserIDs$, $Tweets$ a set of tweets

```
userID ← next(UserIDs)
userCluster ← next(UserClusters)
tweet ← next(Tweets)
while Tweets not empty do
    if tweet.userID < userID then
        tweet ← next(Tweets)
    else if tweet.userID == userID then
        cluster ← openfile(userCluster, "append")
        cluster.write(tweet)
        cluster.close()
        tweet ← next(Tweets)
    else
        userID ← next(UserIDs)
        userCluster ← next(UserClusters)
    end if
end while
```

4.3 Sentiment Analysis

We chose the lexicon-based SentiStrength tool [89] to perform automatic sentiment classification of the tweets, because (1) it does not require training data when working in unsupervised mode; (2) it has good performance and it is able to process more than $16,000$ tweets/second in machines with 64 bit CPU of 3.33 Ghz, and 16.0 GB RAM [86] – similar to the machine used in our work; (3) and has good results on Twitter datasets [89, 38].

It is possible to obtain a free license of Java version of SentiStrength for research purposes, although usually a commercial version is used. For that we contacted the authors, requesting a free license for the most recent Java version of SentiStrength and we also asked for configuration data for English, Spanish, Portuguese and German. They gave us the Java version of SentiStrength and the files needed to use it for those four languages, warning us that the English files were the most tested and reliable version, while the Spanish, Portuguese and German files were adaptations made by several students, not exhaustively tested.

This program is a Java JAR application **SentiStrengthCom.jar** that needs the following files to do the Sentiment Classification:

- BoosterWordList.txt – List of words that increase the strength of the immediately following words.
- EmoticonLookupTable.txt – Table of emoticons' sentiment values.
- EmotionLookupTable.txt – Table of words' sentiment values.
- IdiomLookupTable.txt – Table of idioms' sentiment values.
- IronyTerms.txt – List of ironic words.
- NegatingWordList.txt – List of negating words.
- QuestionWords.txt – List of question words.

These files contain all the linguistic information needed for the sentiment classification. Using SentiStrength for different languages implies adapting these files to the characteristics of each language.

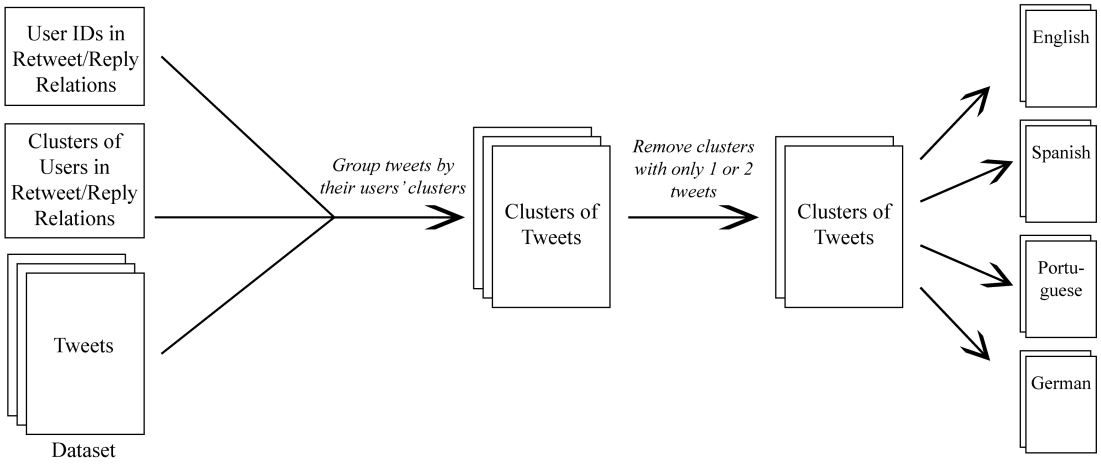


Figure 4.3: Tweet clustering process.

SentiStrength offers different options of execution which allow the adaptation of the strength and priority of certain rules in the algorithm, but we chose to run it in the default modus operandi.

SentiStrength receives Giving a text file as input and outputs another file with each line of text of the input file annotated with two sentiment values: a positive integer $s_+ \in \{1, ..., 5\}$ and a negative integer $s_- \in \{-5, ..., -1\}$. The higher the absolute value, the higher the polarity strength of that value.

Input:

```
My respect to the team from #CostaRica! #WorldCup2014
Congrats Argentina!!!! #WorldCup2014
Congrats to our guy Messi #teamadidas #allin or nothing
Don't mess with Messi
```

Output:

```
Positive        Negative        Text
3    -1  My respect to the team from #CostaRica! #WorldCup2014
2    -1  Congrats Argentina!!!! #WorldCup2014
2    -1  Congrats to our guy Messi #teamadidas #allin or nothing
1    -1  Don't mess with Messi
```

Before we performed any systematic testing on SentiStrength results, we noticed two contrasting situations regarding the quality of the linguistic files. The Spanish files were a solid adaptation of the original ones, having even more terms and idioms than the English version. On the contrary, the Portuguese files, despite having a similar number of terms when compared with the English version, lead the algorithm to classify a larger number of texts with the neutral value $(1, -1)$ than it would be expected. The problem was in the adaptation of the file **EmotionLookupTable.txt** which had a large table of Portuguese words and respective sentiment values, but it was not used the star notation to cover a broader number of terms. The star notation is used when there are many different word endings for a word, and they share an equal sentiment meaning. This only applies to word endings, in which the word is truncated and the end of the word is replaced with the star "*". For instance, "love*" would match "love", "lover", and "loved" so that separate entries are not needed for all of these terms. However, it is important

to not use the star notation in words that could end in a suffix with a completely different sentiment from other possible suffixes.

We fixed this file by analyzing each word and its possible suffixes, truncating and adding the star notation when we only found similar sentiment meaning among the possible suffixes. This task was not performed by an expert, and it is possible that some words have been wrongly annotated with the star notation. We also introduced some idioms in the Portuguese version of **IdiomLookupTable.txt**, especially related with football. No modifications were made for English, Spanish or German.

To classify the tweets in each cluster of tweets we filtered only the tweet text. To avoid words out of context that could be matched by SentiStrength, we removed all the mentions, retweet indicatives and URLs occurrences in the text. Hashtags were not removed because they are usually related with the context of the text. We used the following regular expressions for matching these occurrences:

```
1 Mentions and Retweet indicative matching: (RT \)*@[[:alnum:]_]*[: ]*
2 URL matching: http[s]*\:\/\/[[:alnum:]=?#&\/_\-\.]*
```

We could have used the **"entities"** object in the original JSON structure of the tweet to do this filtering, however parsing the JSON structure would be more computationally demanding.

After running SentiStrength over the clusters of tweets we got for each cluster a matching file with the classified sentiment annotated for each tweet.

4.4 Influence and Sentiment Homophily Analysis over Time

The user clustering, tweet clustering and sentiment analysis stages were scripted to extract the information about the clusters in the network and their sentiment, during desired time intervals.

We observe sentiment changes over time, but we do not know how often it changes, so we exploited three different time intervals for clustering:

1. Global clustering, from the oldest tie in the dataset on March 13th until the last one on July 15th.
2. Daily-based clustering, using the knock-out stage subset, from June 27th until July 15th.
3. Round-based clustering, using the knock-out stage subset, considering the period of the round of 16, quarter-finals, semi-finals and final stages of the World Cup.

With these different types of clusters we aimed to understand which time range gives the best clustering results for our dataset and which gives more insight about the overall sentiment of the clusters. However, we used each clustering for different purposes, for which we defined different metrics and methods. With daily-based clustering we tested the hypothesis of narrow time clusters evidence an overall sentiment homophily. With round-base clustering we assumed that sentiment homophily is highly dynamic and we searched for it locally in time.

4.4.1 Sentiment Homophily in Narrow Time Clusters

This strategy pursues our first objective of testing if there is a clusters' prevalent sentiment. For that we built the network and clustered it according to the relations that occurred in a narrow period of time. Then, we assumed that the sentiment inside these clusters as a static property. The purpose of the daily-based clustering was to find out if clusters show a homogeneous sentiment in ranges of 24 hours.

We analyzed this by simply calculating the sentiment distribution for each file, considering the sentiment of each tweet in the cluster. Once we have both positive and negative values of sentiment strength, we calculated three different distributions for:

Absolute Sentiment value,

$$|s| = s_+ + s_-, \in \{-4, ..., 0, ..., 4\} \tag{4.2}$$

Independent Sentiment value,

$$s_{ind} = s_+ \cup s_-, \in \{-5, ..., -1, 1, ..., 5\} \tag{4.3}$$

Sentiment Pair value,

$$s_{pair} = (s_+, s-), \in \{(1, -1), ...{}_{24}, (5, -5)\} \tag{4.4}$$

These distributions were calculated for each cluster, in each language, in each day, independently. We also included information about the presence of mentions, URLs and hashtags, and we combined the results in an overall distribution of all clusters, for each day, in each language.

4.4.2 Polarity Changes in Wide Time Clusters

Following our second objective, we look for the sentiment dynamics during the lifetime of each cluster. We focused on polarity changes over time. The purpose of the round-based clustering, as well as the global clustering, was creating a time window for sentiment changes in each cluster. We analyzed the overall sentiment of these clusters in ranges of one hour.

Here we used a more strict approach regarding the way we interpreted the sentiment results. Since we were seeking an overall sentiment, we chose to condensate the two sentiment values in one unique value, calculating the Absolute Sentiment value. This way, a tweet is positive with a strength between 1 and 4, neutral when 0, or negative with a strength between -1 and -4. This approach promotes clearly polarized sentiment results and penalizes balanced strength results. This way, the results $(5, -5)$, $(4, -4)$, $(3, -3)$, $(2, -2)$, which we consider ambiguous, have the same absolute sentiment of 0 as the SentiStrength neutral result $(1, -1)$.

We calculated the distribution of the absolute sentiment values per hour, in each cluster, and we decided to empower the strength of the classification, giving each sentiment result a weight equal to its absolute sentiment value.

These distributions were calculated by counting the number of tweets with each absolute sentiment result, considering its respective weight, for each hour since the hour of the first tweet until the hour of the last tweet in the cluster. The set of results is $\{-4, -3, -2, -1, 1, 2, 3, 4, \}$, and the absolute value 0 was excluded because its weight is 0. The result was stored in the format "Comma-separated values" as follows, which we used to plot the results using the visualization D3.js library [4]:

```
1 hour,-4,-3,-2,-1,1,2,3,4
2 2014-07-08 06:00:00,0,0,0,0,0,0,2,0,8
3 2014-07-08 07:00:00,0,0,0,0,0,0,0,0,20
4 2014-07-08 08:00:00,0,0,0,0,0,0,0,0,8
5 2014-07-08 09:00:00,0,0,0,0,0,0,0,0,0
6 2014-07-08 10:00:00,0,0,0,0,0,0,0,0,0
7 2014-07-08 11:00:00,0,0,0,0,0,0,0,0,0
8 2014-07-08 12:00:00,0,0,0,0,0,0,0,0,0
```

By time-lining the sentiment dynamics in a one hour scale, we expected to identify polarity spikes of sentiment and understand how long they last.

[4] http://d3js.org/

4.4.3 Local Sentiment Homophily in Wide Time Clusters

For the third goal of this work, we propose a technique for finding moments of prevalent polarized sentiment which we used to understand whether a prevalent sentiment can be extrapolated to the rest of the cluster, or not. If there is evidence of this hypothesis being true, the same technique can be used to improve ambiguous classifications of sentiment, using the overall sentiment of the cluster.

To systematically find periods of polarity homophily, assuming that sentiment homophily is found locally in time, we defined a time window t, a minimum number of tweets m needed to consider a sentiment prevalence in t, and minimum rate of polarity prevalence p in t, as metric for sentiment homogeneity. Let $\Delta t(x_1, x_2)$ be the time interval between two tweets, and $pol(x_1, ..., x_n)$ be the rate of the prevalent polarity in a sequence of tweets, there is sentiment homophily for a sequence of tweets $x_1, x_2, ..., x_n$ when,

$$n \geq m \wedge pol(x_1, ..., x_n) \geq p \wedge \forall \{x_i, x_{i+1}, ..., x_{i+m}\} \in \{x_1, x_2, ..., x_n\}, \Delta t(x_i, x_{i+m}) \leq t. \quad (4.5)$$

However, finding time intervals that satisfy this metric does not show if there is an increased chance of any user in that cluster of sharing the same befitting sentiment with the overall sentiment that surrounds him, i.e., being influenced by his peers' mood. Our approach to evaluate whether moments of sentiment homophily are caused by influence is to look for ambiguous tweets in moments of prevalent polarized sentiment in the cluster, to which we assign that same prevalent polarization, and then we compare this updated sentiment classification with human coder classifications. To evaluate the extent of sentiment homogeneity in those periods we used K-fold Cross Validation [106].

Lets assume the pairs $(1, -1), (2, -2), (3, -3), (4, -4), (5, -5)$ as ambiguous results in polarized clusters. The reason for this assumption regarding $(2, -2), (3, -3), (4, -4), (5, -5)$ is that they reveal sentiment strength but not a decided polarization, even in a polarized environment. We also include $(1, -1)$ because SentiStrength outputs this value both for neutral sentences and for sentences that do not match any word in the lexicon, which gives a dubious meaning to this value. This way, we trust more in polarized pairs.

After identifying ambiguous results, we search for an ambiguity a that has a number of surrounding tweets equal or greater than m, with a prevalence of a certain polarity equal or greater than p during a period of time t that includes a. For each ambiguity a, found in a context with these characteristics, we set its polarity to be the same as the prevalent polarity of the tweets surrounding it. We proposed two Algorithms 3 and 4, that only differ in the position that the ambiguity occupies in the context configuration.

The first Algorithm 3 searches for ambiguities that have a central position in the polarized context, being fixed at the center of the time window. For a set of ambiguities A found in a sequence of tweets $T = \{x_1, ..., x_n\}$, when $x_a \in A \wedge x_a \in T$, and

$$\exists x_b, x_e \in T, (b \leq a < e \vee b < a \leq e) \wedge \Delta t(x_b, x_a) \leq \frac{t}{2} \wedge \Delta t(x_a, x_e) \leq \frac{t}{2} \wedge e - b \geq m \wedge pol(x_b, x_e) \geq p, \quad (4.6)$$

The sentiment polarity of x_a is relabeled with the prevalent sentiment polarity in $x_b, ..., x_e$.

The second Algorithm 4 considers any ambiguity that belongs to a sliding time window t that fulfills those restrictions, independently of its position in the context. For a set of ambiguities A found in a sequence of tweets $T = \{x_1, ..., x_n\}$, when $x_a \in A \wedge x_a \in T$, and

$$\exists x_b, x_e \in T, (b \leq a < e \vee b < a \leq e) \wedge \Delta t(x_b, x_e) \leq t \wedge e - b \geq m \wedge pol(x_b, x_e) \geq p, \quad (4.7)$$

The sentiment polarity of x_a is relabeled with the prevalent sentiment polarity in $x_b, ..., x_e$.

This technique is used to reason about the extent of influence in moments of sentiment homophily in clusters, but it is also possible that social selection or an external source, such a natural disaster

may cause moments of sentiment homophily. To understand better the role of social influence in these moments, we can adapt the Algorithm 3 to search for ambiguities at the beginning of t and to search for ambiguities at the end of t, instead of searching for central ambiguities. This way we can compare the two results and, if there is a higher chance of ambiguities to have a befitting sentiment with their clusters when they are at the end of t rather when they are at the beginning of t, then it could indicate that homophily is caused by social influence, because it means that homophily tends to happens latter in time.

Algorithm 3 Ambiguities Sentiment Context – Fixed time window

Require: t time window in hours, $context_{min}$ minimum number of tweets acceptable in t, p prevalent sentiment rate in t, $Cluster$ a set of tweets with sentiment classification sorted by date

$date_{list} \leftarrow []$
$sentiment_{list} \leftarrow []$
$ambiguity_{list} \leftarrow []$
$i \leftarrow 0$
for t in $Cluster$ **do**
    $date_{list}$.append($t.date$) ▷ Date in seconds
    $sentiment_{abs} \leftarrow t.sentiment_{+} + t.sentiment_{-}$
    $sentiment_{list}$.append($sentiment_{abs}$)
    if $sentiment_{abs} == 0$ **then**
        $ambiguity_{list}$.append(i)
    end if
    $i \leftarrow i + 1$
end for
for a in $ambiguity_{list}$ **do**
    $i_{from} \leftarrow a$
    $i_{to} \leftarrow a$
    $positive_{count} \leftarrow 0$
    $negative_{count} \leftarrow 0$
    while $(i_{from} >= 0)$ and $(date_{list}[i_{from}] >= (date_{list}[a] - (t/2) * 3600))$ **do**
        if $sentiment_{list}[i_{from}] > 0$ **then**
            $positive_{count} \leftarrow positive_{count} + 1$
        end if
        if $sentiment_{list}[i_{from}] < 0$ **then**
            $negative_{count} \leftarrow negative_{count} + 1$
        end if
        $i_{from} \leftarrow i_{from} - 1$
    end while
    while $(i_{to} < length(Cluster))$ and $(date_{list}[i_{to}] <= (date_{list}[a] + (t/2) * 3600))$ **do**
        if $sentiment_{list}[i_{to}] > 0$ **then**
            $positive_{count} \leftarrow positive_{count} + 1$
        end if
        if $sentiment_{list}[i_{to}] < 0$ **then**
            $negative_{count} \leftarrow negative_{count} + 1$
        end if
        $i_{to} \leftarrow i_{to} + 1$
    end while
    $i_{from} \leftarrow i_{from} + 1$ ▷ Fix last iteration
    $i_{to} \leftarrow i_{to} - 1$ ▷ Fix last iteration
    $context_{size} \leftarrow (i_{to} - i_{from}) + 1$
    if $(context_{size} >= context_{min})$ and $((positive_{count}/context_{size}) >= p)$ **then**
        print(Surrounding of a: positive)
    end if
    if $(context_{size} >= context_{min})$ and $((negative_{count}/context_{size}) >= p)$ **then**
        print(Surrounding of a: negative)
    end if
end for

Algorithm 4 Ambiguities Sentiment Context – Sliding time window

Require: t time window in hours, $context_{min}$ minimum number of tweets acceptable in t, p prevalent sentiment rate in t, $Cluster$ a set of tweets with sentiment classification sorted by date

$date_{list} \leftarrow []$
$sentiment_{list} \leftarrow []$
$ambiguity_{list} \leftarrow []$
$i \leftarrow 0$
for t in $Cluster$ **do**
    $date_{list}$.append($t.date$) ▷ Date in seconds
    $sentiment_{abs} \leftarrow t.sentiment_{+} + t.sentiment_{-}$
    $sentiment_{list}$.append($sentiment_{abs}$)
    if $sentiment_{abs} == 0$ **then**
        $ambiguity_{list}$.append(i)
    end if
    $i \leftarrow i + 1$
end for
for a in $ambiguity_{list}$ **do**
    $i_{from} \leftarrow a$
    $i_{to} \leftarrow a$
    $positive_{count} \leftarrow 0$
    $negative_{count} \leftarrow 0$
    while (not $(i_{from} <= 0)$ and $(i_{to} >= (length(Cluster) - 1))$) and $((date_{list}[i_{to}] - date_{list}[i_{from}]) <= t * 3600)$ **do**
        $\Delta t_{before} \leftarrow MAXINT$
        $\Delta t_{after} \leftarrow MAXINT$
        if $i_{from} > 0$ **then**
            $\Delta t_{before} = date_{list}[i_{from}] - date_{list}[i_{from} - 1]$
        end if
        if $i_{to} < (length(Cluster) - 1)$ **then**
            $\Delta t_{after} = date_{list}[i_{to} + 1] - date_{list}[i_{to}]$
        end if
        if $(\Delta t_{before} == \Delta t_{after})$ or $(\Delta t_{before} < \Delta t_{after})$ **then**
            $i_{from} \leftarrow i_{from} - 1$
            if $sentiment_{list}[i_{from}] > 0$ **then**
                $positive_{count} \leftarrow positive_{count} + 1$
            end if
            if $sentiment_{list}[i_{from}] < 0$ **then**
                $negative_{count} \leftarrow negative_{count} + 1$
            end if
        else
            $i_{to} \leftarrow i_{to} + 1$
            if $sentiment_{list}[i_{from}] > 0$ **then**
                $positive_{count} \leftarrow positive_{count} + 1$
            end if
            if $sentiment_{list}[i_{from}] < 0$ **then**
                $negative_{count} \leftarrow negative_{count} + 1$
            end if
        end if
    end while
    $context_{size} \leftarrow (i_{to} - i_{from}) + 1$
    if $(context_{size} >= context_{min})$ and $((positive_{count}/context_{size}) >= p)$ **then**
        print(Surrounding of a: positive)
    end if
    if $(context_{size} >= context_{min})$ and $((negative_{count}/context_{size}) >= p)$ **then**
        print(Surrounding of a: negative)
    end if
end for

Chapter 5

Results

In this chapter we present and discuss the results obtained from the different stages of our work, described in Chapter 4. Since we used three different periods of time for clustering, we first present a comparative description of the clustering results, for each time interval. Then, we describe the results of the Influence and Homophily analysis considering each method and respective clustering set used. More results can be found at `http://link.inesc-id.pt/sentiment/`.

5.1 User Clustering

With global clustering we got social circles based on relations that happened over 4 months. As it is represented in Table 5.1, $24,987,618$ different users were involved in retweet relations that sum up a total of $115,820,265$ different relations among them, while $10,490,130$ users were connected through $18,944,754$ reply relations. The number of retweet relations is then 4 times higher than the number of users involved, while the number of reply relations is less than 2 times the users involved in reply relations. This not only shows that retweets are more frequent than replies, it also means that people tend to relate with a more restrictive spectrum of other users in reply relations than they do in retweet interactions.

The clustering results also point that replies tend to be more restrict, originating smaller and denser clusters than retweets, like the cluster size distribution in Figure 5.1) suggests. Both types of interactions exhibit a power-law regarding the cluster size distribution. The exponent of the retweet-based set is 1.87512, while in the reply-based set is 2.02154, i.e., the frequency of clusters decreases more abruptly with the size of the cluster for replies, having also a higher frequency of small clusters than retweets. The size of the largest cluster resulting from retweets covers more than half of all nodes, while in replies it only covers about 15%.

By analyzing the connected components of each graph we observed that a giant component arises in both cases and they cover the majority of the clusters, except a portion of the smaller ones.

Global Clustering	***RT***	***RE***
Nodes	24,987,618	10,490,130
Edges – Directed Graph	115,820,265	18,944,754
Edges – Symmetric Graph	229,766,965	34,774,682
Clusters	494,488	1,772,577
Larger Cluster Size	14,757,442	1,666,870
Giant Component Size	24,403,970	8,382,804

Table 5.1: Summary of Global User Clustering characteristics.

Daily Clustering - RT	*Nodes*	*Edges - Directed Graph*	*Edges - Symmetric Graph*	*Clusters*	*Larger Cluster Size*	*Giant Component Size*	*Games*
June 27th	1,110,327	1,380,120	2,753,940	108,567	20,777	940,648	–
June 28th	1,752,054	2,793,320	5,564,746	152,866	45,586	1,575,209	BRA-CHI COL-URU
June 29th	1,603,921	2,354,072	4,688,204	150,112	25,767	1,423,640	NED-MEX CRC-GRE
June 30th	1,568,063	2,288,270	4,553,970	137,936	55,015	1,391,965	FRA-NGA GER-ALG
July 1st	2,097,046	3,112,379	6,200,926	205,468	16,634	1,865,358	ARG-SUI BEL-USA
July 2nd	1,404,633	1,809,588	3,602,512	125,536	21,631	1,224,865	–
July 3rd	1,010,800	1,274,790	2,534,272	95,846	10,647	857,601	–
July 4th	1,864,501	3,078,044	6,131,644	160,428	25,431	1,691,522	BRA-COL FRA-GER
July 5th	2,328,379	4,259,272	8,489,736	158,646	265,216	2,163,756	NED-CRC ARG-BEL
July 6th	1,277,434	1,784,684	3,552,274	105,907	19,595	1,135,584	–
July 7th	1,166,225	1,543,893	3,072,146	101,875	21,534	1,016,019	–
July 8th	2,063,306	2,974,598	5,931,320	207,415	85,446	1,812,637	BRA-GER
July 9th	2,367,939	3,604,730	7,184,860	205,917	33,359	2,137,234	NED-ARG
July 10th	1,674,548	2,613,922	5,216,896	138,703	18,855	1,493,823	–
July 11th	1,301,290	1,780,165	3,554,222	117,087	20,298	1,137,131	–
July 12th	1,404,797	2,125,146	4,242,307	120,746	13,330	1,255,081	BRA-NED
July 13th	**2,513,013**	**3,862,576**	**7,711,354**	**250,646**	**28,048**	**2,241,499**	GER-ARG
July 14th	2,188,244	3,375,638	6,741,768	179,237	37,403	1,973,061	–
July 15th	1,173,111	1,556,139	3,107,530	105,747	36,711	1,010,382	–

Table 5.2: Summary of Daily-based User Clustering characteristics per day for **retweets** graph, with information about the games schedule.

The Daily-based User Clustering gives an insight about the dynamics of relations over time for our topic-centric dataset, as it is stated in Table 5.2 and Table 5.3. We can notice that there is a burst of nodes and relations in game days, and that burst varies with the game itself. As we can see in the day of the Final, on July 13th, registers the highest number of users in the graph.

The differences between the clusters obtained from the graphs of retweets and replies with the global clustering are also noticeable in the daily-based clustering, but with even more expression for the rate between edges and nodes, and the rate between giant component size and nodes that reveal a stronger decrease for replies. However, contrarily to the global clustering, the amount of clusters obtained with the retweets graph was higher than the amount obtained with the replies graph.

In what regards the cluster size distribution, retweets originated power-laws with exponent between 1.87949 and 1.94033, while replies originated power-laws with exponent between 2.15865 and 2.45473. This shows that the frequency of clusters decreases more abruptly with the size of the cluster for replies, like we also have seen for the global clustering. These distributions revealed a similar order of magnitude, and Figure 5.2 and Figure 5.3 represent the distributions of the day with the median value of power-law exponent, respectively, for retweets and replies.

With the round-based user clustering we obtained similar results to the global clustering, regarding the number of nodes, edges, clusters, and cluster size. The major difference is in the size of the giant component for replies is considerably smaller than the total number of nodes, as Table 5.4 and Table 5.5 show. The clusters that are not contained in the giant component are mainly clusters of small size as we can see in Figure 5.4. The cluster size distribution of all four stages reveals identical characteristics.

Daily Clustering - RE	*Nodes*	*Edges - Directed Graph*	*Edges - Symmetric Graph*	*Clusters*	*Larger Cluster Size*	*Giant Component Size*	*Games*
June 27th	247,932	194,516	365,489	83,434	1,710	75,779	–
June 28th	369,923	302,789	566,624	120,040	2,404	131,038	BRA-CHI COL-URU
June 29th	376,552	306,818	577,160	120,259	6,875	129,133	NED-MEX CRC-GRE
June 30th	357,128	294,245	551,095	114,936	4,814	130,821	FRA-NGA GER-ALG
July 1st	458,044	372,248	697,206	149,778	2,601	161,231	ARG-SUI BEL-USA
July 2nd	275,823	217,723	408,372	91,362	1,372	88,107	–
July 3rd	221,955	174,341	326,873	74,509	1,155	67,958	–
July 4th	398,074	328,564	616,181	127,735	4,307	146,990	BRA-COL FRA-GER
July 5th	479,095	434,753	806,853	142,164	8,279	217,115	NED-CRC ARG-BEL
July 6th	249,511	204,359	383,325	78,982	4,529	90,164	–
July 7th	230,940	184,945	347,823	74,940	2,144	76,730	–
July 8th	408,990	316,540	597,437	140,320	3,020	121,056	BRA-GER
July 9th	484,440	397,837	740,845	159,108	7,565	168,672	NED-ARG
July 10th	335,955	276,389	517,963	105,382	18,918	124,226	–
July 11th	287,887	235,279	439,040	92,892	3,799	100,255	–
July 12th	298,179	244,082	456,105	96,884	5,352	104,400	BRA-NED
July 13th	**524,817**	**428,565**	**809,048**	**168,361**	**8,063**	**190,539**	GER-ARG
July 14th	416,985	362,972	677,618	123,494	19,485	179,251	–
July 15th	249,886	200,143	375,737	80,581	9,939	85,219	–

Table 5.3: Summary of Daily-based User Clustering characteristics per day for **replies** graph, with information about the games schedule.

Round Clustering - RT	*Nodes*	*Edges - Directed Graph*	*Edges - Symmetric Graph*	*Clusters*	*Larger Cluster Size*	*Giant Component Size*
Round of 16	5,500,239	11,165,748	22,231,862	291,702	2,287,154	5,157,046
Quarter-finals	4,758,745	9,890,395	19,694,802	229,830	1,030,843	4,472,301
Semi-finals	4,888,105	9,392,322	18,707,894	279,271	1,518,466	4,555,534
Final	6,303,393	14,223,580	28,372,313	305,115	1,572,557	5,931,270

Table 5.4: Summary of Round-based User Clustering characteristics for **retweets**.

Round Clustering - RE	*Nodes*	*Edges - Directed Graph*	*Edges - Symmetric Graph*	*Clusters*	*Larger Cluster Size*	*Giant Component Size*
Round of 16	1,440,841	1,427,570	2,664,758	401,017	15,977	755,738
Quarter-finals	1,144,461	1,126,502	2,096,587	320,343	10,602	590,013
Semi-finals	1,145,818	1,075,240	2,008,503	334,578	13,352	542,494
Final	1,630,666	1,681,422	3,136,273	435,474	54,931	889,730

Table 5.5: Summary of Round-based User Clustering characteristics for **replies**.

With these results we found that wider time intervals for clustering give more uniform values, while narrower time intervals denote bigger fluctuations depending the events that they cover. Since Global User Clustering and the Round-based clustering revealed similar characteristics, we decided to only use one of them, and we chose Round-based clusters. The reason for that is because they are divided into four comparable rounds which are easier to deal, and also because they comprise less days for analysis.

For Influence and Homophily analysis we must study clusters with several users, so, as we observed with the three types of clustering, we should remove single-node clusters and it could be useful to only consider clusters that belong to the giant component of the graph, because it essentially removes some clusters with small number of users.

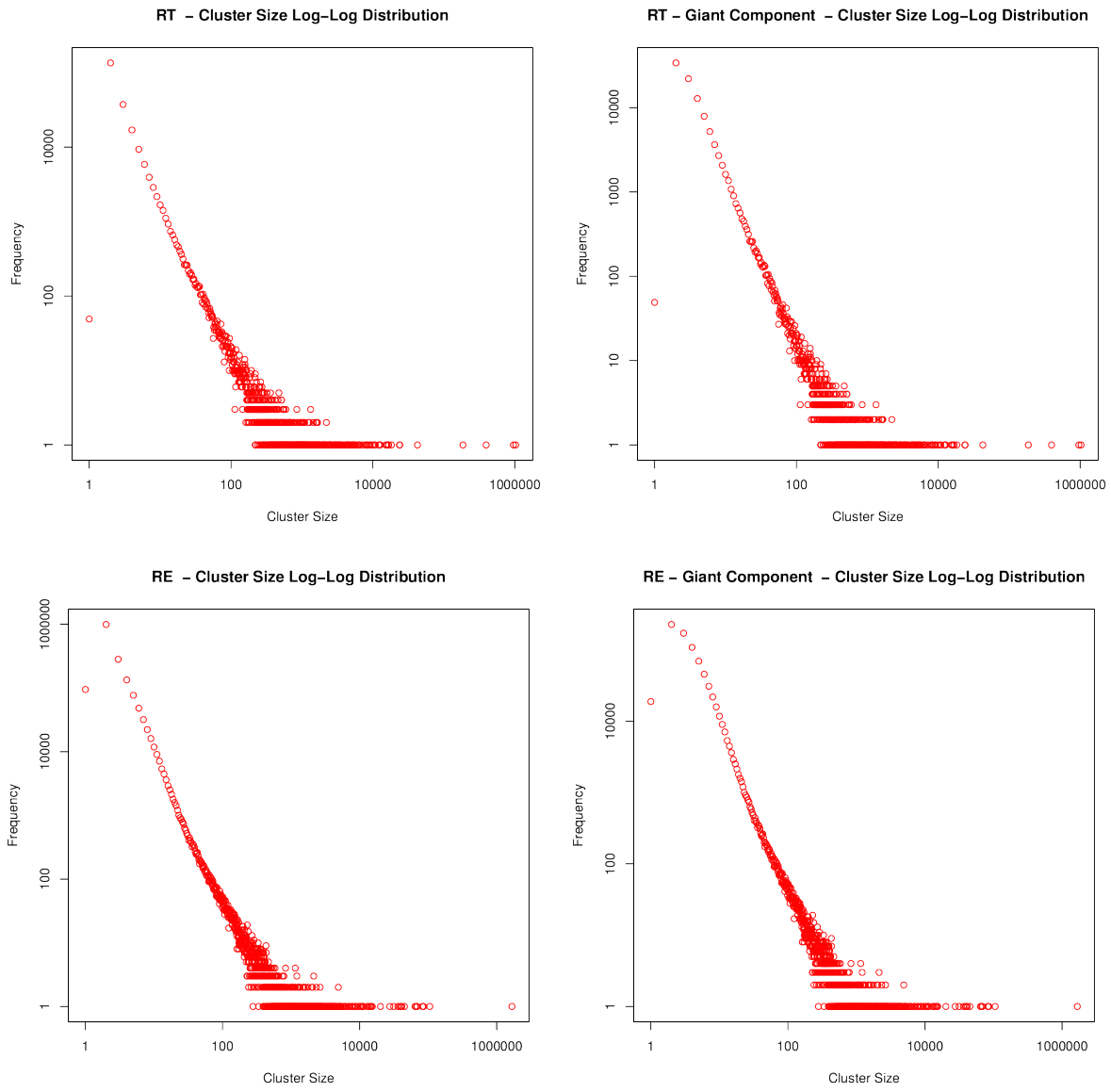


Figure 5.1: Distribution of clusters, obtained from Global User Clustering, regarding their size. On top, the comparison between the totality of **retweet-based** clusters distribution and the distribution of those who belong to the giant component. On bottom, the same comparison for replies. The exponent value of each power-law is, respectively, 1.87512, 1.86919, 2.02154, and 2.02124.

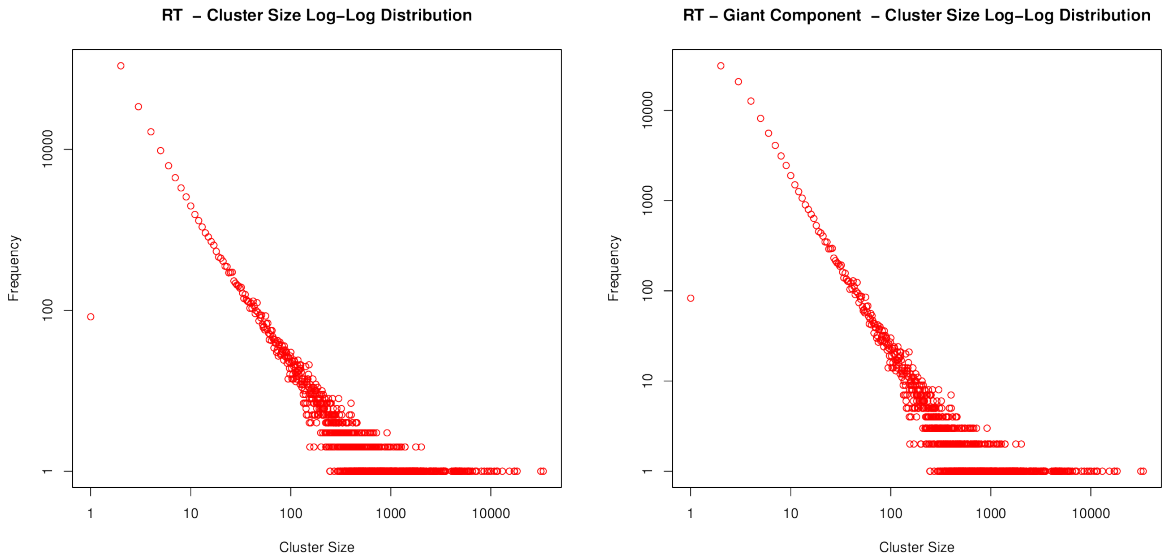


Figure 5.2: Distribution of clusters, obtained from Daily-based Clustering for July 9th, regarding their size. Comparison between the totality of **retweet-based** clusters distribution and the distribution of those who belong to the giant component. The exponent value of each power-law is, respectively, 1.90412 and 1.90028.

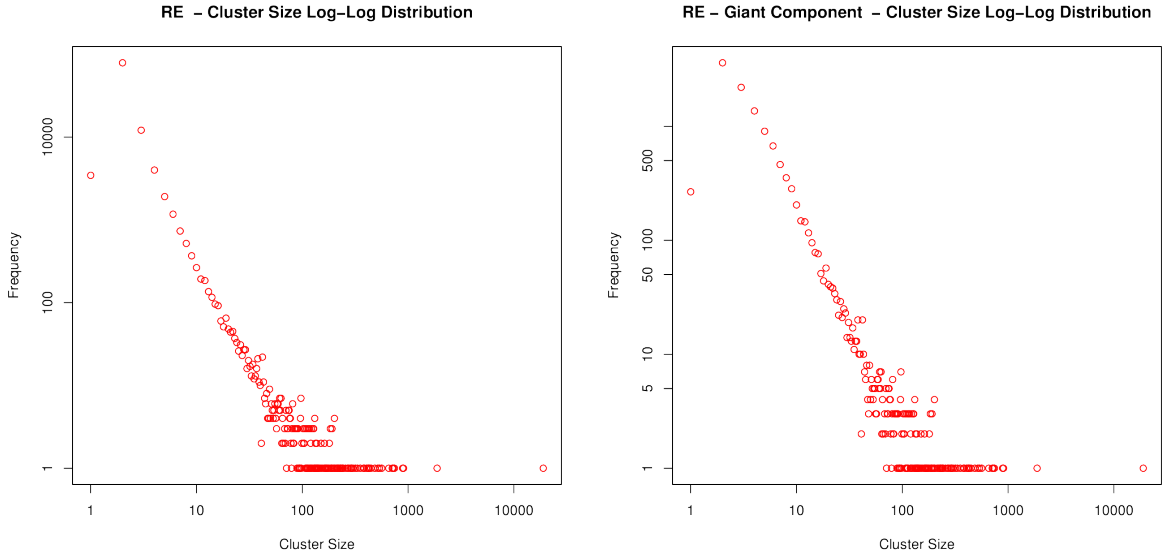


Figure 5.3: Distribution of clusters, obtained from Daily-based Clustering for July 10th, regarding their size. Comparison between the totality of **reply-based** clusters distribution and the distribution of those who belong to the giant component. The exponent value of each power-law is, respectively, 2.21169 and 2.18112.

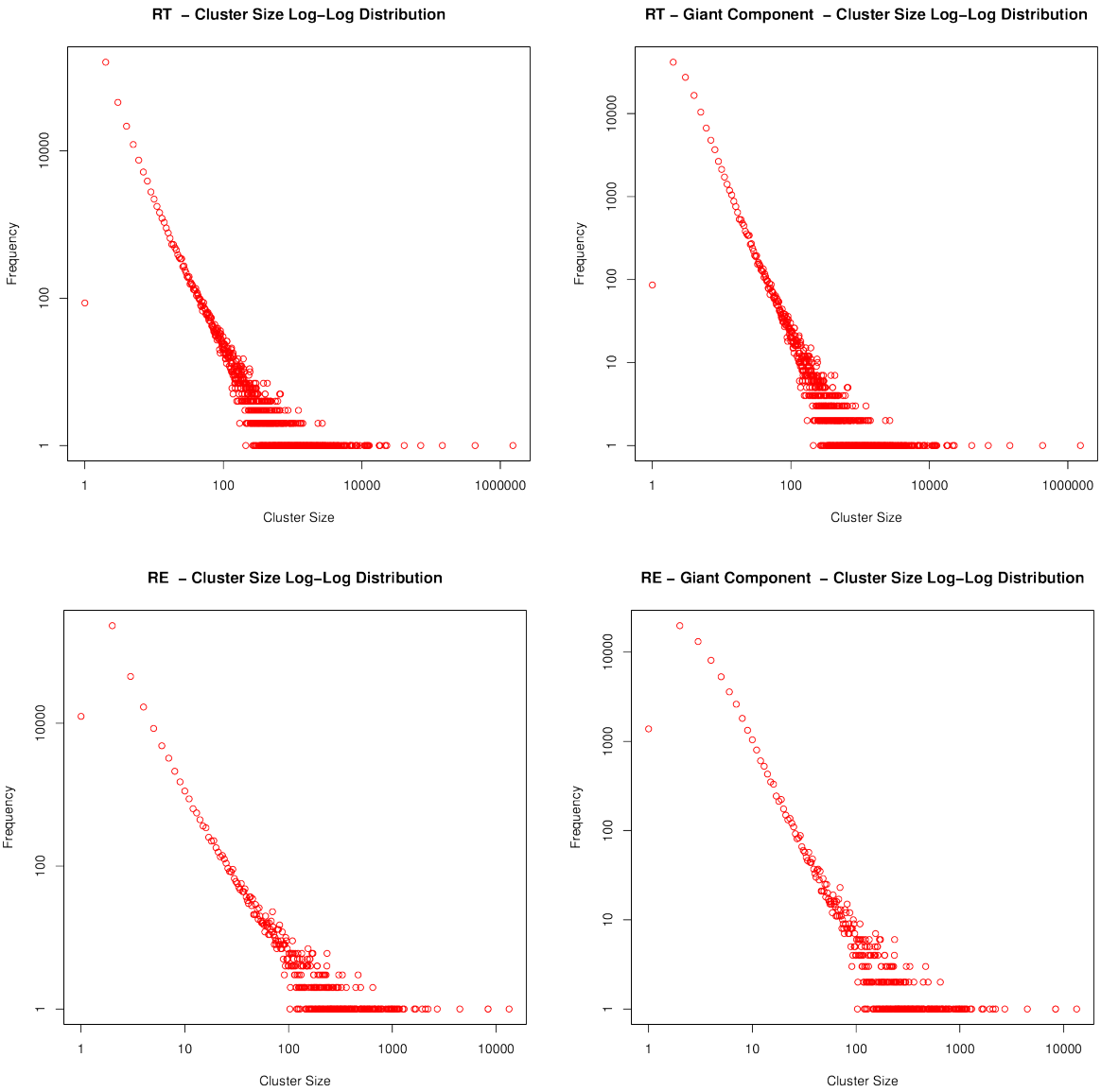


Figure 5.4: Distribution of clusters, obtained from Round-based User Clustering for the Semi-finals, regarding their size. On top, the comparison between the totality of retweet-based clusters distribution and the distribution of those who belong to the giant component. On bottom, the same comparison for replies. The exponent value of each power-law is, respectively, 1.89057, 1.88808, 2.15327, and 2.14271.

Daily-based Clustering	*RT*						*RE*					
	Clusters of Users	***Clusters of Tweets***	***en***	***es***	***pt***	***de***	***Clusters of Users***	***Clusters of Tweets***	***en***	***es***	***pt***	***de***
June 27th	108,567	31,739	24,659	5,557	1,302	221	83,434	20,910	16,866	3,232	746	66
June 28th	152,866	62,472	38,474	13,491	10,294	213	120,040	48,208	33,479	8,898	5,718	113
June 29th	150,112	58,886	43,989	11,897	2,733	267	120,259	44,927	35,987	7,215	1,593	132
June 30th	137,936	48,136	36,827	7,907	2,268	1,134	114,936	39,017	31,430	5,153	1,601	833
July 1st	205,468	93,612	69,542	20,787	2,811	472	149,778	66,287	54,343	9,837	1,873	234
July 2nd	125,536	39,276	31,437	6,075	1,578	186	91,362	26,470	21,982	3,467	946	75
July 3rd	95,846	27,644	21,061	4,838	1,585	160	74,509	17,810	14,294	2,574	890	52
July 4th	160,428	56,832	33,252	12,845	10,210	525	127,735	42,838	28,779	7,848	5,900	311
July 5th	158,646	62,205	36,829	21,515	3,578	283	142,164	59,888	36,535	14,701	8,455	197
July 6th	105,907	31,876	19,813	7,793	4,142	128	78,982	20,736	14,243	4,054	2,375	64
July 7th	101,875	28,886	19,665	5,751	3,304	166	74,940	17,586	12,835	2,812	1,869	70
July 8th	207,415	67,669	46,842	11,757	8,417	653	140,320	44,001	33,000	5,316	5,442	243
July 9th	205,917	77,726	49,528	20,430	7,342	426	159,108	57,563	42,357	10,179	4,869	158
July 10th	138,703	48,533	26,871	18,056	3,382	224	105,382	31,319	19,436	9,517	2,283	83
July 11th	117,087	39,814	24,880	12,165	2,574	195	92,892	25,489	17,707	6,154	1,551	77
July 12th	120,746	44,302	28,209	9,732	6,079	282	96,884	32,016	22,954	5,003	3,939	120
July 13th	250,646	96,387	64,373	24,695	6,307	10,12	168,361	66,306	50,650	11,550	3,692	414
July 14th	179,237	61,129	38,657	18,397	3,304	771	123,494	40,548	26,155	11,952	2,133	308
July 15th	105,747	31,128	21,730	7,521	1,503	374	80,581	20,350	14,434	4,900	862	154
Total	2,828,685	1,008,252	676,638	241,209	82,713	7,692	2,145,161	722,269	527,466	134,362	56,737	3,704
%	**100%**	**35.6%**	**67.1%**	**23.9%**	**8.2%**	**0.8%**	**100%**	**33.7%**	**73.0%**	**18.6%**	**7.9%**	**0.5%**

Table 5.6: Comparison between the number of clusters of users and the number of clusters of tweets, obtained with daily-based clustering. The differences between retweets and replies in each language are also compared.

Round-based Clustering	*RT*						*RE*					
	Clusters of Users	***Clusters of Tweets***	***en***	***es***	***pt***	***de***	***Clusters of Users***	***Clusters of Tweets***	***en***	***es***	***pt***	***de***
Round of 16	291,702	123,058	96,011	21,365	4,879	803	401,017	213,795	163,133	36,960	12,418	1,284
Quarter-finals	229,830	81,609	64,071	13,032	3,850	656	320,343	151,082	102,941	31,322	16,177	642
Semi-finals	279,271	102,418	74,873	22,058	4,736	751	334,578	150,587	108,042	26,215	15,743	587
Final	305,115	114,454	98,705	11,155	3,584	1,010	435,474	219,925	153,295	50,811	14,660	1,159
Total	1,105,918	421,539	333,660	67,610	17,049	3,220	1,491,412	735,389	527,411	145,308	58,998	3,672
%	**100%**	**38.1%**	**79.2%**	**16.0%**	**4.0%**	**0.8%**	**100%**	**49.3%**	**71.7%**	**19.8%**	**8.0%**	**0.5%**

Table 5.7: Comparison between the number of clusters of users and the number of clusters of tweets, obtained with round-based clustering. The differences between retweets and replies in each language are also compared.

5.2 Tweet Clustering

By clustering the users of each graph we got clusters with different numbers of users. To study Influence and Homophily mechanisms, we require clusters with multiple users. For that we removed single-node clusters and looked only for the giant component of each graph. To get the sentiment of each cluster we clustered users' tweets by their respective cluster and we divided them by language. Once again we assumed that clusters with only 1 or 2 tweets give a poor insight of the sentiment of the cluster, and they were also removed from our analysis.

The deletion of small clusters and language filtering reduced the number of clusters in more than half the number of the initial clusters, as we can see in Tables 5.6 and 5.7. This reduction seems to be lower with wider time range for clustering. In what concerns the number of tweets per cluster, we concluded that the cluster size distribution for tweets has the same nature as the distribution for users, exhibiting a power-law. Regardless of the type of relation of the graph, the time period of the clustering, or the language, we found a large number of small clusters followed by a small number of big clusters. Furthermore, we observed a tendency for having one or two particularly big clusters, as we can see for example in Figures 5.5, 5.6, 5.7, and 5.8 are examples of that.

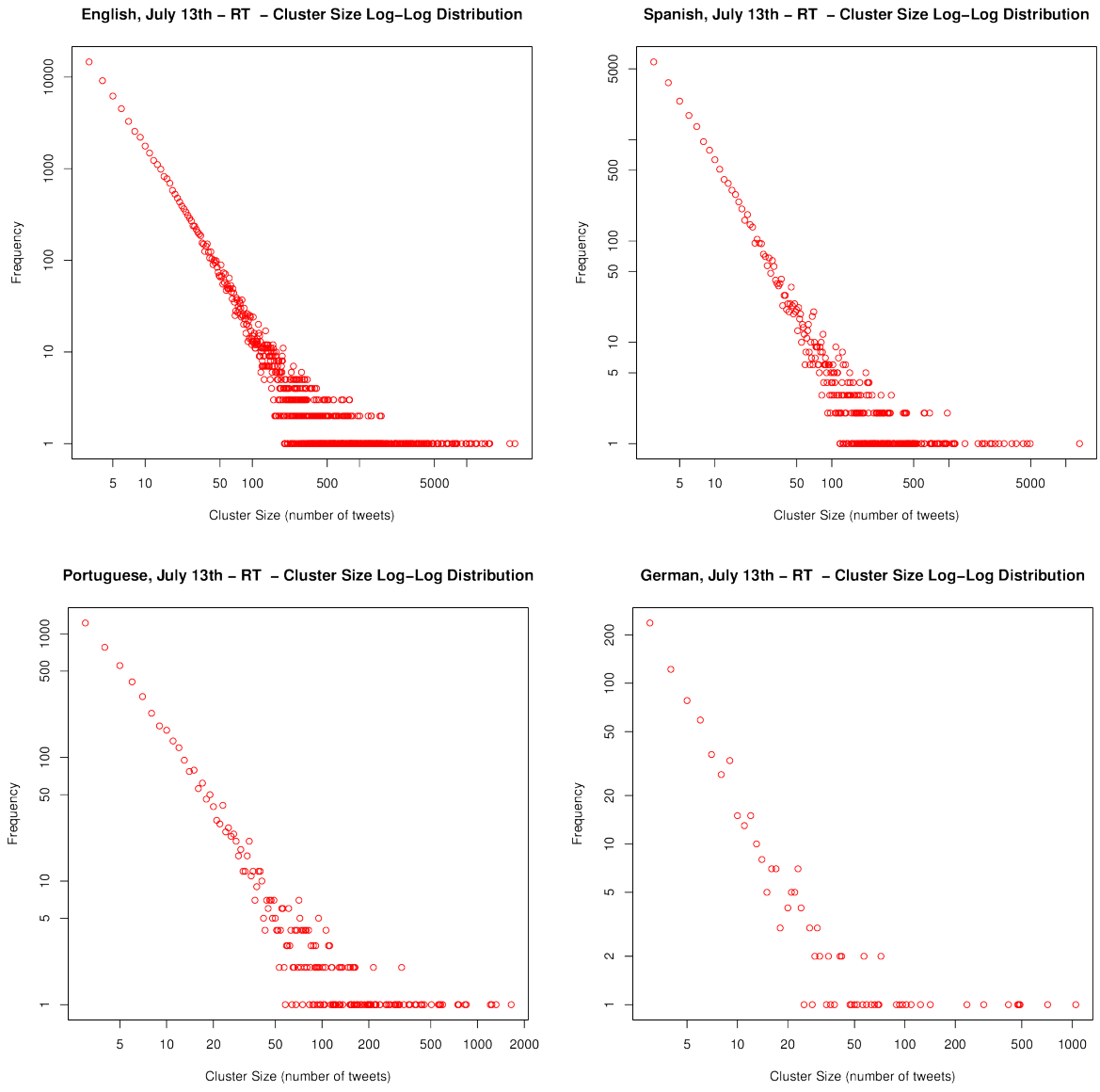


Figure 5.5: Distribution of clusters, obtained from Daily-based Tweet Clustering for **retweets** on July 13th, regarding their number of tweets.

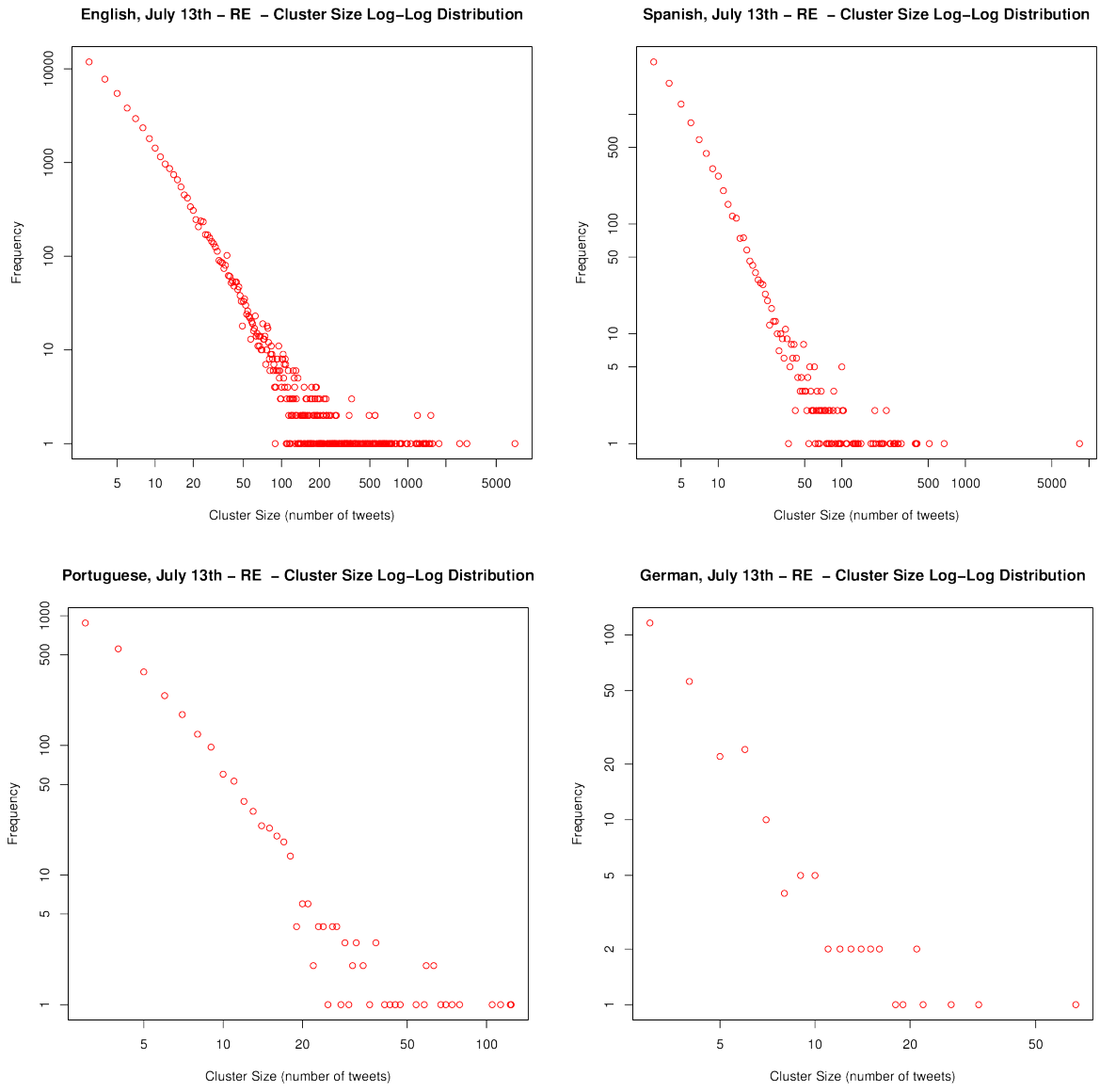


Figure 5.6: Distribution of clusters, obtained from Daily-based Tweet Clustering for **replies** on July 13th, regarding their number of tweets.

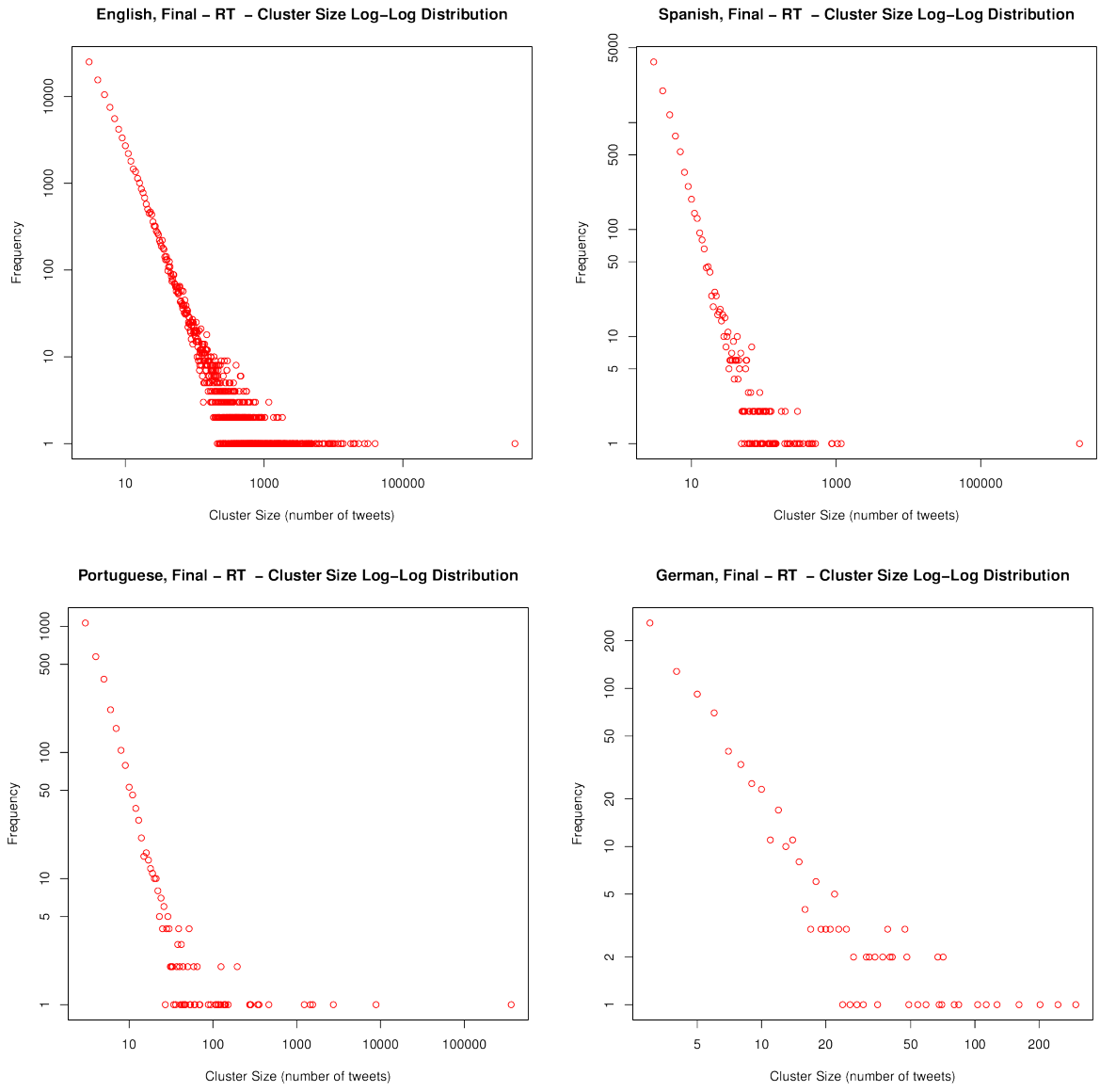


Figure 5.7: Distribution of clusters, obtained from Round-based Tweet Clustering for **retweets** on the Final stage, regarding their number of tweets.

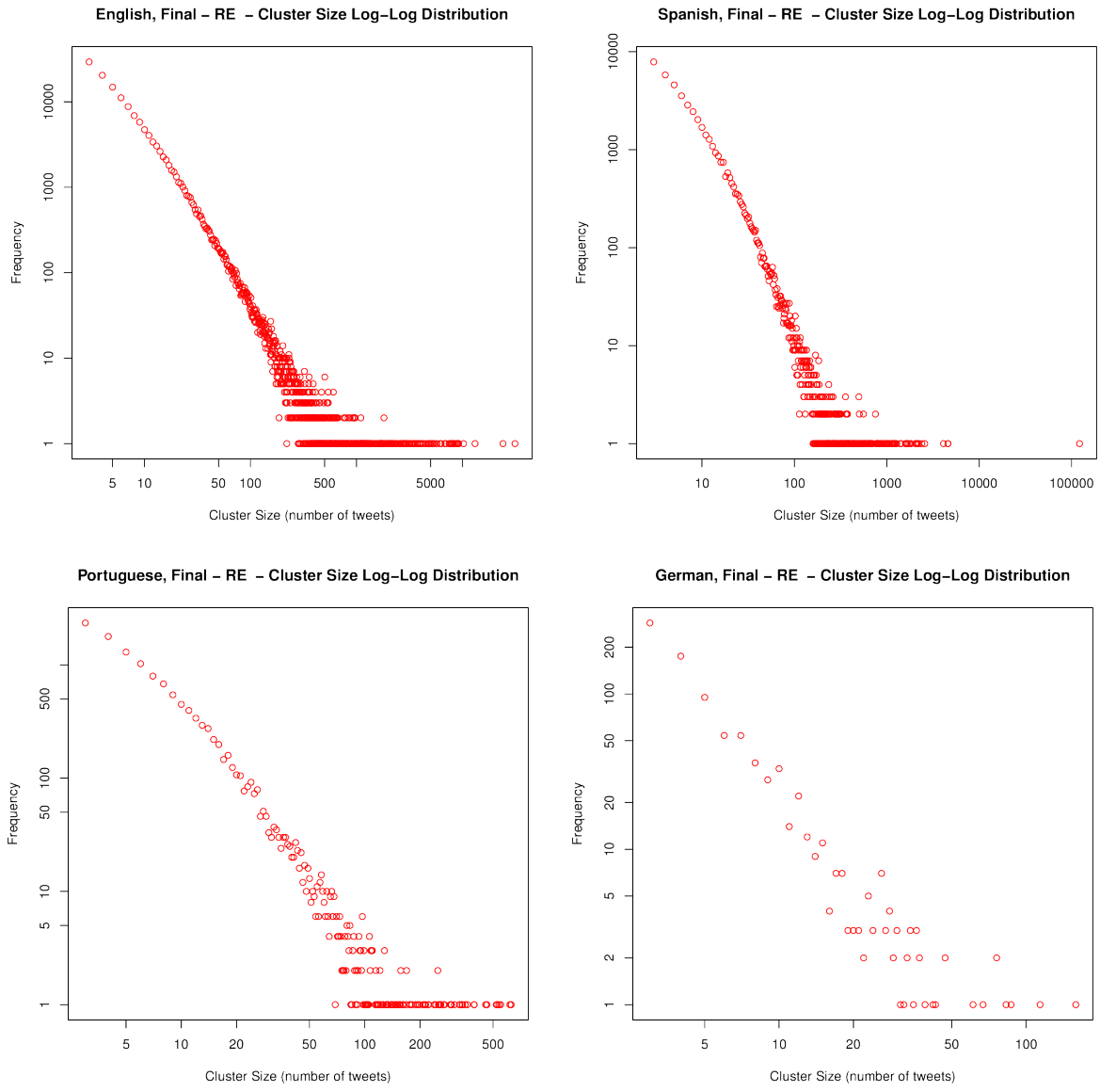


Figure 5.8: Distribution of clusters, obtained from Round-based Tweet Clustering for **replies** on the Final stage, regarding their number of tweets.

5.3 Influence and Sentiment Homophily Analysis over Time

The results above describe the clusters used for our analysis Influence and Sentiment Homophily analysis. Those clusters are divided by relation type, time period and language, and we chose these divisions because they allow us to replicate our experiments in similar but independent sets of clusters with particular different characteristics, such as the matches they cover. All tweets in these clusters of tweets have a sentiment classification. By comparing their sentiment results we may find a correlation with some features or universal patterns among these different groups.

5.3.1 Sentiment Homophily in Narrow Time Clusters

The purpose of Daily-based clustering was to test a narrow interval of time that could be used to obtain good clustering results and, at the same time, cover the minimum period of time necessary to have sentiment homogeneity. We set the period of 24 hours as the time window to look for sentiment homophily, and we simply analyzed the sentiment distribution present in each cluster file, as a whole. In this analysis we considered three different interpretations of the sentiment classification: the absolute sentiment value, independent sentiment value, and sentiment pair value.

Our first finding was that the independent sentiment value, by separating the positive value from the negative value, highlight repetition but may hide polarization. For instance, the set of results $(5, -1)$, $(3, -1)$, $(4, -1)$ has three positive results, but this approach will enhance the repetition of -1. Actually, -1 was the most frequent result in this distribution over the different sets of clusters. Since this approach causes the loss of information about polarization we discarded it.

On the other hand, both the absolute sentiment value and the sentiment pair value keep the information about polarization and the major conclusion that we found in the configuration of their distributions is that the period of 24 hours does not give evidence of sentiment homophily. Despite the fact that the absolute value of 0 and the pair $(1, -1)$ are the values with higher frequency over the majority of the clusters, they usually appear alongside with polarized values. We found homogeneous sentiment results mostly over clusters of small size, what is intrinsically related to the fact that they have few tweets and, consequently, small variety of sentiment classifications.

We found that the bigger the size of the cluster, the higher the tendency for sentiment heterogeneity even with prevalence of the absolute value 0 and the pair $(1, -1)$. Figures 5.9 and Figure 5.10 are two examples that show this prevailing characteristic, independently of the language, the day, and relation type.

To better understand the extension of this apparently repetitive result, we joined the clusters per day, type of relation, and language and we calculated their overall distribution. As we can notice in Figure 5.11, the absolute values 0 and 1, and the pairs $(1, -1)$ and $(2, -1)$ were always the two most frequent results. We can also observe that the difference between each consecutive absolute value is more gradual for English, and the Portuguese and Spanish clusters contain more occurrences of stronger values of sentiment, while the German clusters reveal a neutral configuration. These differences may be related with different criteria used in each language adaptation of SentiStrength files. It is important to stress that we analyzed these overall distributions one-by-one and, regardless of the day, this pattern is found repeatedly for both retweet and reply-based clusters.

With this strategy we understood that if there is Sentiment Homophily in a cluster, a 24 hours window is too large to notice its evidence. However, according to the distribution of sentiment values, clusters seem to be neutral most of the time, with episodes of polarity changes that we should find locally in time.

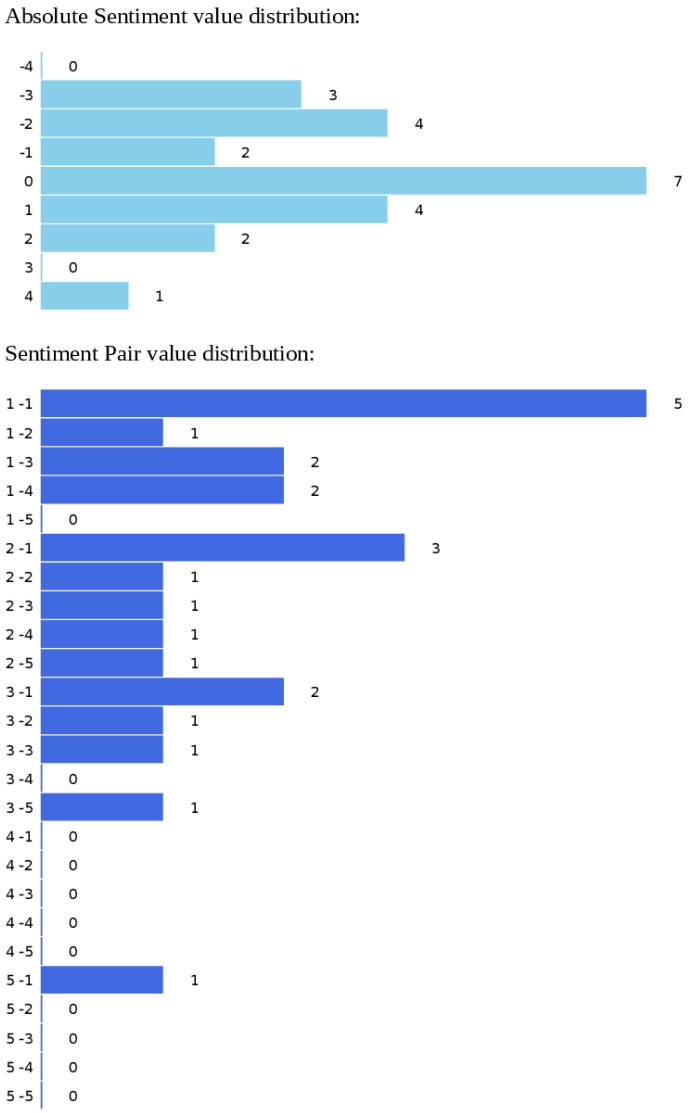


Figure 5.9: Distribution of absolute sentiment values and sentiment pair values in Spanish-speaking cluster “334972” on July 13th. Cluster size: 23 tweets.

5.3.2 Polarity Changes in Wide Time Clusters

Using a static time window turned out to not be an appropriate technique to find sentiment homogeneity in social circles, but it unveiled that different sentiment values usually arise during the clusters’ lifetime, denoting a general prevalence for neutrality and a weak positive tendency in our dataset.

To better understand the sentiment dynamics on Twitter, we looked for polarity changes in longer time intervals. We used the round-based clusters to have longer periods of clusters’ lifetime to analyze, and we measured the frequency of each possible absolute sentiment value for each hour in each cluster. With this time-line we looked for an informal evidence of sentiment homophily.

We observed that the majority of the clusters contain polarity spikes over their lifetime, as we can see in Table 5.8. The set of clusters with larger expression of completely neutral clusters are the German-speaking clusters obtained from the retweet-relation graph, during the Round of 16, that are 36.2% of the total number of clusters. Moreover, German-speaking clusters, followed by Portuguese-speaking clusters, were the sets with higher percentage of neutral clusters, while English and Spanish-speaking clusters revealed similar lower rates of neutrality. The quality of language configuration files for Sen-

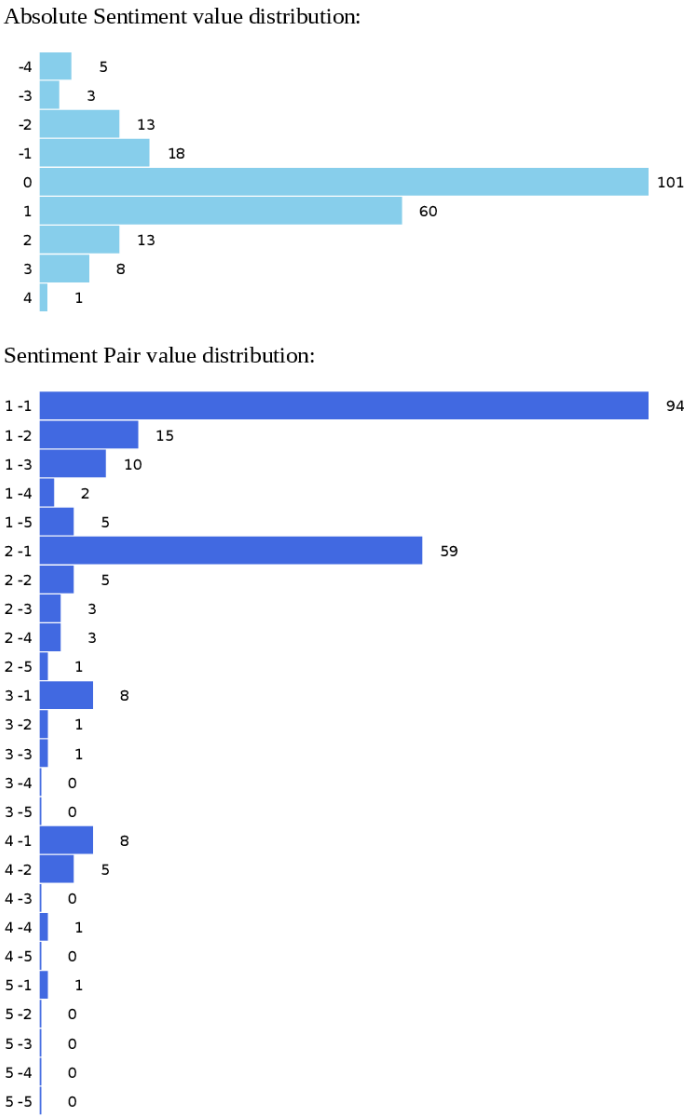


Figure 5.10: Distribution of absolute sentiment values and sentiment pair values in Portuguese-speaking cluster “797328” on July 13th. Cluster size: 222 tweets.

tiStrength may be the reason for these differences, because the lexicon size for English and Spanish is substantially bigger than the lexicon size of Portuguese and German files. One should remember that SentiStrength classifies sentences that do not match words in the lexicon as neutral.

Independently of the language or stage of the championship we observed that sets of reply-based clusters tend to have fewer neutral clusters than the retweet-based clusters. Since reply-based clusters include conversations among their tweets, this could indicate that conversational tweets tend to be less neutral.

To observe how sentiment changes over time and detect sentiment homophily, we plotted the timeline of tweet frequency according to the corresponding absolute sentiment value, Figure5.12. We noticed that polarized sentiment arises in spikes and it is considerably volatile. These moments of polarization appear at different rate and with different frequency values, but they are usually ephemeral, lasting just a few hours, regardless of the set of clusters considered.

Tweets may be dispersed over time, originating lack of sentiment information in certain intervals of time, and lower sentiment frequencies. When this happens on small size clusters the amount of sen-

Stage	*Language*	*Type*	*Clusters with polarity spikes*	*Neutral clusters*
Round of 16	*en*	*RT*	86,929 (90.5%)	9,082 (09.5%)
		RE	158,856 (97.4%)	4,277 (02.6%)
	es	*RT*	19,594 (91.7%)	1,771 (08.3%)
		RE	35,554 (96.2%)	1,406 (03.8%)
	pt	*RT*	3,680 (75.4%)	1,199 (24.6%)
		RE	11,220 (90.4%)	1,198 (09.6%)
	de	*RT*	512 (63.8%)	291 (36.2%)
		RE	1,130 (88.0%)	154 (12.0%)
Quarter-finals	*en*	*RT*	57,578 (89.9%)	6,493 (10.1%)
		RE	100,057 (97.2%)	2,884 (02.8%)
	es	*RT*	11,588 (88.9%)	1,444 (11.1%)
		RE	30,081 (96.0%)	1,241 (04.0%)
	pt	*RT*	2,935 (76.2%)	915 (23.8%)
		RE	15,025 (92.9%)	1,152 (07.1%)
	de	*RT*	487 (74.2%)	169 (25.8%)
		RE	561 (87.4%)	81 (12.6%)
Semi-finals	*en*	*RT*	67,542 (90.2%)	7,331 (09.8%)
		RE	104,827 (97.0%)	3,215 (03.0%)
	es	*RT*	19,901 (90.2%)	2,157 (09.8%)
		RE	25,005 (95.4%)	1,210 (04.6%)
	pt	*RT*	3,626 (76.6%)	1,110 (23.4%)
		RE	14,525 (92.3%)	1,218 (07.7%)
	de	*RT*	511 (68.0%)	240 (32.0%)
		RE	482 (82.1%)	105 (17.9%)
Final	*en*	*RT*	89,033 (90.2%)	9,672 (09.8%)
		RE	148,525 (96.9%)	4,770 (03.1%)
	es	*RT*	9,765 (87.5%)	1,390 (12.5%)
		RE	49,156 (98.7%)	1,655 (03.3%)
	pt	*RT*	2,819 (78.7%)	765 (21.3%)
		RE	13,423 (91.6%)	1,237 (08.4%)
	de	*RT*	650 (64.4%)	360 (35.6%)
		RE	957 (82.6%)	202 (17.4%)

Table 5.8: Number of completely neutral clusters and the number of clusters with polarity spikes for round-based clusters.

timent data for each hour may be insufficient to detect sentiment homophily. For instance, the cluster in Figure 5.13 show spikes dispersed in time and separated by intervals of hours, which give few information about a possible sentiment prevalence, despite having spikes of polarized sentiment. This is a problem that we found in the majority of German-speaking clusters because of their small number of tweets. These time intervals between sentiment peaks could be moments of tweet absence, but also moments of neutral or ambiguous tweets that would not be plotted because they have weight 0. It is plausible that there could be neutral sentiment homophily, however, as we mentioned on Chapter 4, SentiStrength classifies both neutral texts and mismatching texts with the same value, which biases the meaning of that classification. For this reason we decided to focus our attention on polarity homogeneity, discarding neutrality.

Small clusters give poor sentiment information, while clusters at the end of the clusters size power-law give very heterogeneous results with a large amount of different sentiment classifications. For instance, the cluster in Figure 5.14, which is the largest cluster in its set, shows an equivalent sentiment burst of both positive and negative sentiments. The biggest burst, on July 8th, appears during the Brazil elimination, by loosing with Germany for the score of $1 - 7$. It is possible that this type of event triggers antagonistic reactions, but as this cluster includes a huge number of users, it is also possible that these larger clusters could be divided into other minor clusters for this type of analysis. This phenomenon was found in all sets of clusters for English, Spanish and Portuguese, and Figure 5.15 is another example of that.

Informally, a cluster reveals polarity homophily in a given period of time if, during that time, there is a considerable number of tweets and the majority of them share the same polarity value. Actually, we found different clusters which, at some point of their time-line, suggested the existence of a sentiment homogeneity. Moreover, the cause of this phenomenon seems to be different for retweets and replies. These findings seem to be transversal to the four languages.

In retweet-based clusters, we frequently found sentiment bursts that are clearly distinguishable from the surrounding sentiment. These bursts were mainly retweet chains that created a homogeneous sentiment prevalence in the cluster. If we assume that, when a user retweets a status of another user, there is a higher chance of the retweeter sharing the same sentiment of the original tweet, then we can say that in a retweet chain there is sentiment homophily caused by an influence action. Figures 5.17, and 5.18 illustrate sentiment homophily caused by chains of retweets, during different periods of time.

In the reply-based clusters we found a conversational context that usually originated an entanglement of different sentiments, from which sometimes raised a prevalent polarity caused by some subject-related discussion. Figure 5.19 shows negative homophily arising from an entangled surrounding, while Figure 5.12 shows a similar case for positive homophily. For instance, Figure 5.20 denotes a more ephemeral case of polarity homogeneity in a quieter context.

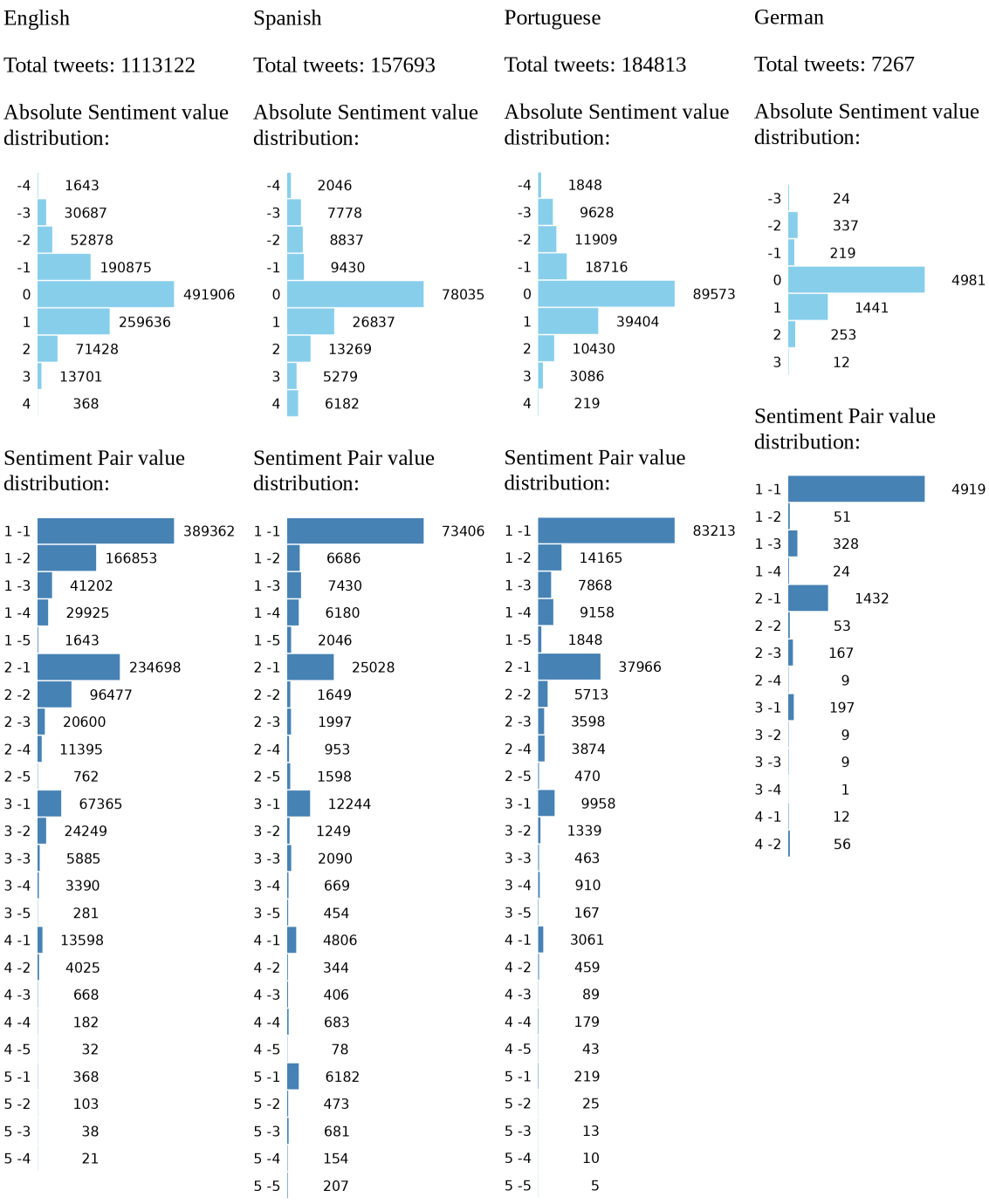


Figure 5.11: Comparison between the overall distribution results for the four languages on July 8th, considering retweet-relation clusters.

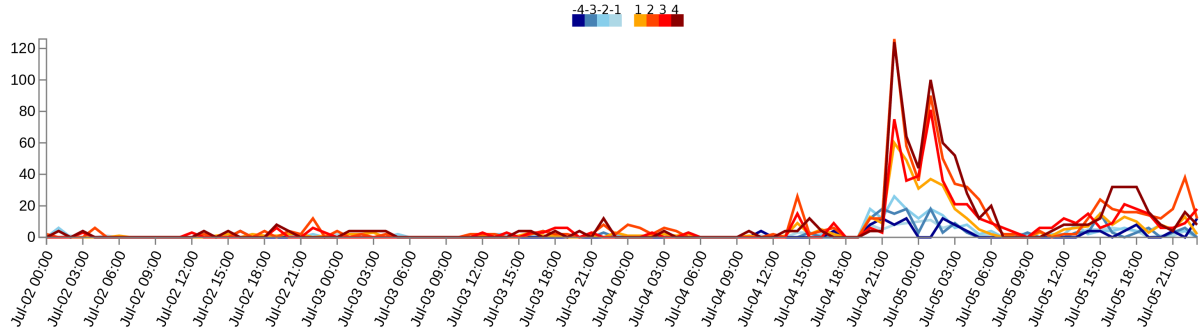


Figure 5.12: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "413547" from the Spanish-speaking set of **reply-based** clusters over the Quarter-finals stage.

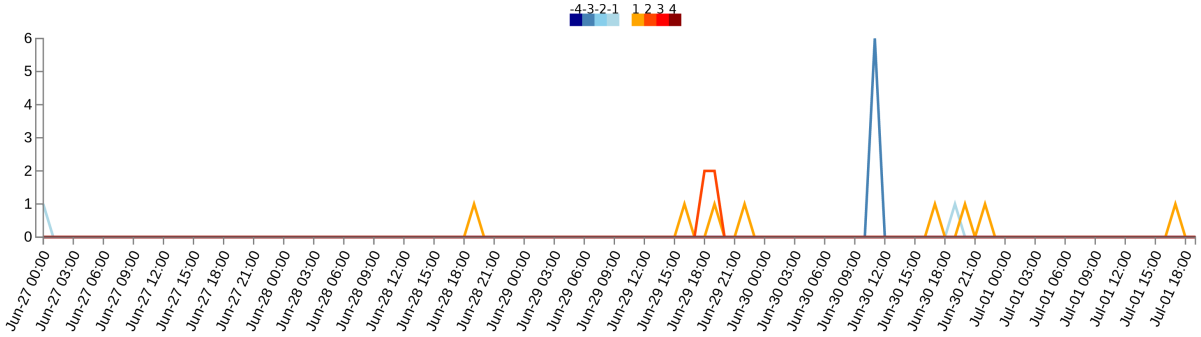


Figure 5.13: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "5171567" from the English-speaking set of **retweet-based** clusters over the Round of 16.

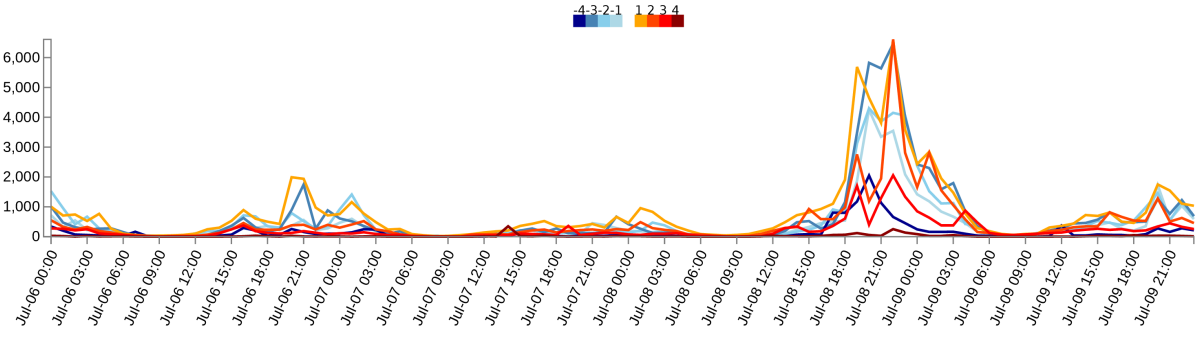


Figure 5.14: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "553712" from the Portuguese-speaking set of **retweet-based** clusters over the Semi-finals stage.

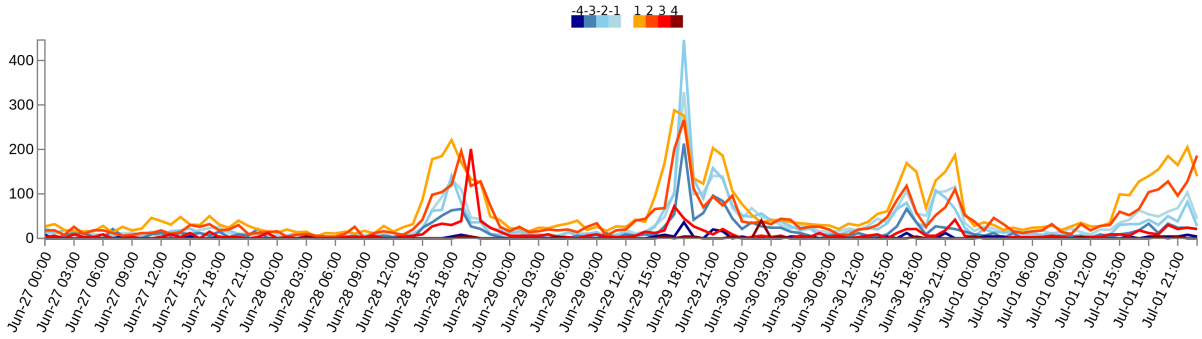


Figure 5.15: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "301837" from the English-speaking set of **reply-based** clusters over the Round of 16.

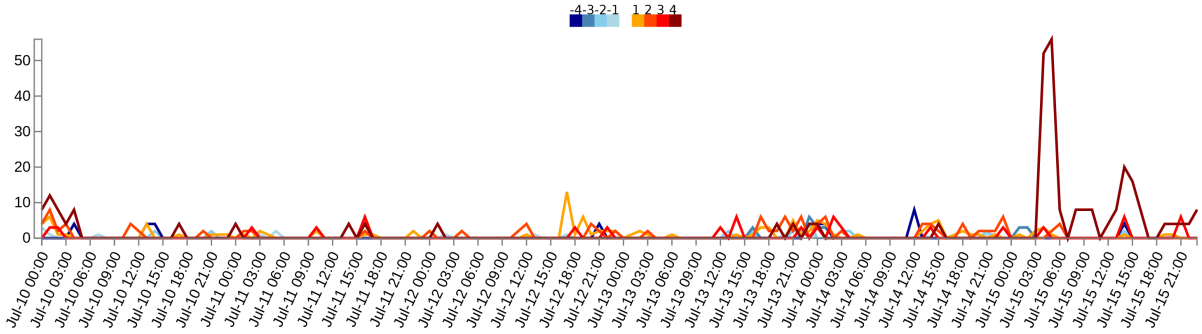
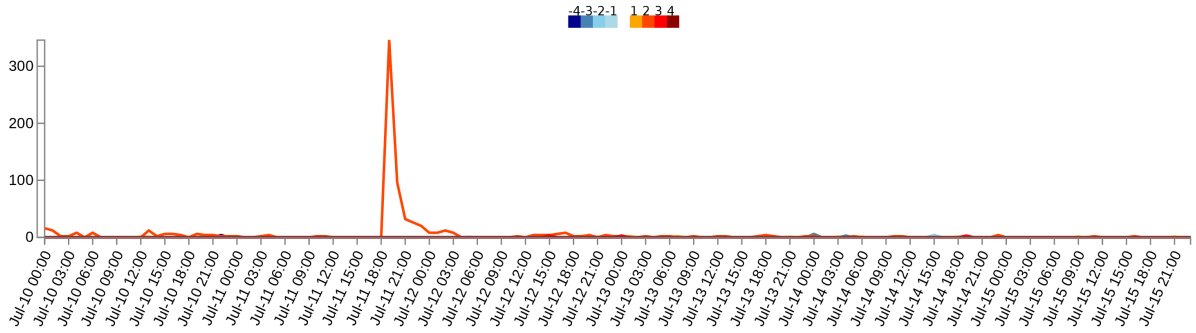


Figure 5.16: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "2049176" from the Spanish-speaking set of **retweet-based** clusters over the Final stage.



Jul-11 18:20 | 3 | -1 | 382613598: Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz.

Jul-11 18:20 | 3 | -1 | 86403361: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 542918347: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 549650191: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 359572574: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 2588021827: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1352273971: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 349298734: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1868335302: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1017273631: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 248906033: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1945438178: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1535394835: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 377652030: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 203107383: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1390825777: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 327015304: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 2342733908: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 2588924911: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 2490588925: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 331300641: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 1689164970: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 928263198: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:20 | 3 | -1 | 502090034: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2170018671: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2355335729: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 203682366: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2408025248: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2411667601: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 160009822: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2480510714: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 257223532: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2259732588: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 1112897982: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 173565593: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 807741314: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 1397188292: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

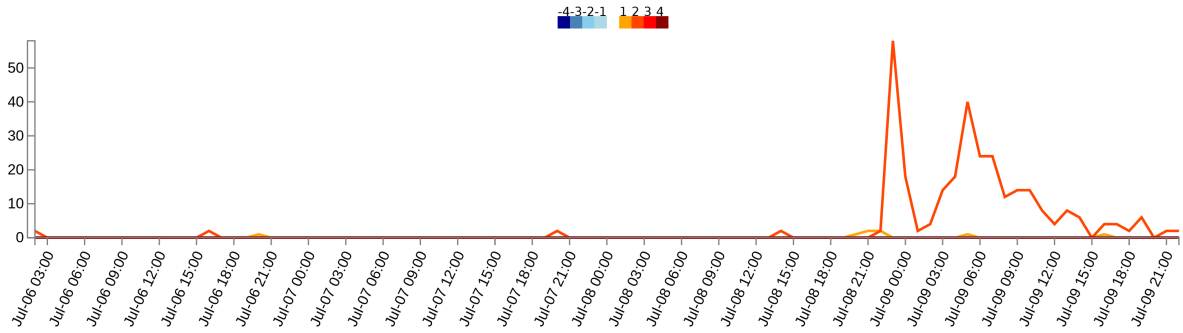
Jul-11 18:21 | 3 | -1 | 319042400: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2502005754: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 1628943684: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Jul-11 18:21 | 3 | -1 | 2630079003: RT @SigaSOJA: "Algum dia, de alguma forma eu vou dar muita alegria pro meu povo - David Luiz."

Figure 5.17: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "2479319" from the Portuguese-speaking set of **retweet-based** clusters over the Final stage.



Jul-08 21:06 | 1 | -1 | 502772808: RT @Br4mm3n: #WorldCup #GERBRA #BRAGER #GER http://t.co/uMgVNSlfy1

Jul-08 21:31 | 3 | -1 | 1978452932: RT @itsMagicNerd : ATEMLOS, DURCH DAS SPIEL! NUR DIE DEUTSCHEN SEHEN DAS ZIEL! #GERBRA #WorldCup2014 #WM2014

Jul-08 21:35 | 2 | -1 | 2512362388: @2014FifaWM mit dem einen Tor oder was?:D Neymar sitzt zuhause und heult bestimmt!

Jul-08 21:40 | 2 | -1 | 59521168: Jaa gegen messi wird echt hart... Bin wirklich gespannt auf Samstag!

Jul-08 21:41 | 1 | -1 | 59521168: Tjaja gegen messi wird echt hart mal gucken was das morgen wird

Jul-08 22:01 | 1 | -1 | 228019862: RT @Br4mm3n: #BRAGER #GER #GERBRA #WorldCup http://t.co/7NLQot569G

Jul-08 22:40 | 3 | -1 | 373024213: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:40 | 3 | -1 | 1000213699: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:40 | 3 | -1 | 2486359521: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 1718804515: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 1958718918: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 741405800: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 403564621: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 2521793500: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 228019862: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:41 | 3 | -1 | 59521168: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:42 | 3 | -1 | 370569429: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:42 | 3 | -1 | 2367484866: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:42 | 3 | -1 | 2257353297: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:43 | 3 | -1 | 1291996014: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:43 | 3 | -1 | 2677124991: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:44 | 3 | -1 | 2665790229: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:46 | 3 | -1 | 2189077738: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:46 | 3 | -1 | 1406135630: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:47 | 3 | -1 | 180883762: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:47 | 3 | -1 | 2198470776: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:48 | 3 | -1 | 961157911: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:49 | 3 | -1 | 1702141129: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:51 | 3 | -1 | 940339010: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:51 | 3 | -1 | 2672075242: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:54 | 3 | -1 | 1222847984: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:55 | 3 | -1 | 2589919214: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:57 | 3 | -1 | 2379084152: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:58 | 3 | -1 | 836047302: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 22:59 | 3 | -1 | 470313353: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:04 | 3 | -1 | 2258514148: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:04 | 3 | -1 | 2170478126: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:05 | 3 | -1 | 28130784: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:08 | 3 | -1 | 89552125: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:11 | 3 | -1 | 2259023841: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Jul-08 23:13 | 3 | -1 | 2397426447: RT @iBlali: Hier, mein Lieblingsfoto von Manuel Neuer. #BRAGER http://t.co/fJPxaTzmIP

Figure 5.18: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "1911770" from the German-speaking set of **retweet-based** clusters over the Semi-finals stage.

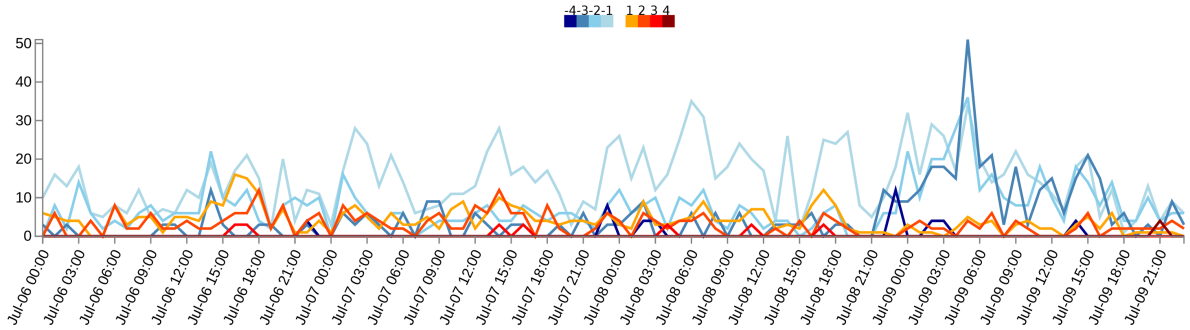


Figure 5.19: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "1000883" from the English-speaking set of **reply-based** clusters over the Semi-finals stage.

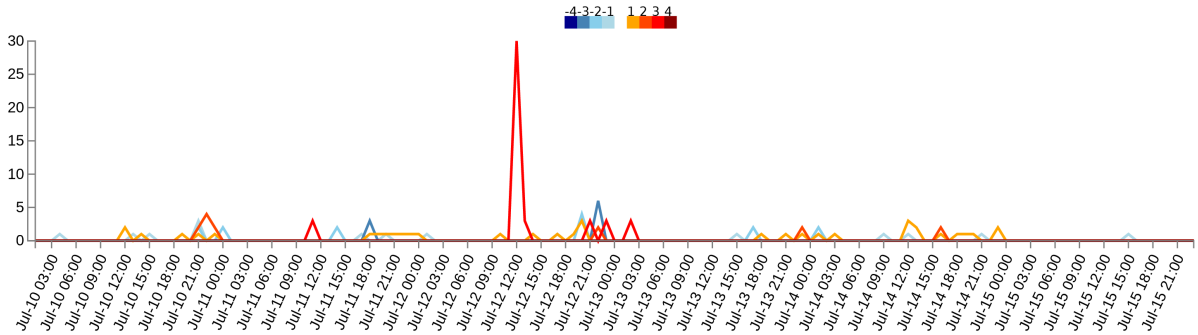


Figure 5.20: Time-line of tweets' frequency of absolute sentiment for each accumulation of 3 hours. Cluster "177613" from the Portuguese-speaking set of **reply-based** clusters over the Final stage.

5.3.3 Local Sentiment Homophily in Wide Time Clusters

Retweet chains and topic-related conversational discussions triggered periods of polarity homophily in several clusters, however this does not happen only associated to these two phenomenons. With the sentiment time-lining we looked for informal evidence of sentiment homogeneity, which we found locally over the clusters' lifetime.

To systematically find periods of polarity homophily we defined a time window t, a minimum number of tweets m needed to consider a sentiment prevalence in t, and minimum rate of polarity prevalence p in t, as metrics for sentiment homogeneity. However, finding time intervals that satisfy this metric does not disclose if there is an increased chance of any user in that cluster of sharing a same befitting sentiment with the overall sentiment surrounding him.

To evaluate this, we searched for dubious sentiment classifications inside sentiment homogeneous periods to which we gave the overall sentiment of their clusters. Then we gave those dubious tweets to human coders in order to understand whether the contextual sentiment increases their probability of fitting in that same sentiment. Although we did not look for neutral homophily, we used neutral absolute sentiment values as dubious classifications, because they include the ambiguous results $(2, -2), (3, -3), (4, -4), (5, -5)$ and the neutral result $(1, -1)$ that is used for both neutral classifications and for unknown sentences.

We used two different strategies for finding ambiguities: Algorithm 3 and Algorithm 4. They differ in the positioning of the time window regarding the ambiguity, in which the first has a fixed time window with the ambiguity in a central position, and the second considers a sliding time window regarding the ambiguity's position. We set $m = 10$ and $p = 0.7$ and compared different time windows (in hours), $t = 1, t = 3, t = 6, t = 12$.

Table 5.9 shows that the number of clusters that contain ambiguities in a polarized context is considerably lower than the number of clusters that have a number of tweets equal or greater than m. However, as we have seen with the clusters' sentiment time-line, tweets are sparse over time and they do not usually create unique spikes of sentiment polarity. So, it is expectable that the homophily constraints will be satisfied in clusters larger than m. The number of clusters obtained with Algorithm 4 is higher than the number obtained with Algorithm 3 because it searches for ambiguities that are included in a homogeneous interval of time, independently of its position in that time interval, which is more flexible than the fixed approach that only searches for ambiguities that are surrounded by homogeneous polarity before and later in time. As long as t grows, the number of clusters obtained converges to the number of clusters with at least 100 tweets.

Figure 5.21 is an example of a tweet with dubious classification that is inside a time interval that meets the conditions of sentiment homophily.

Stage	Language	Type	Cluster size ≥ 10	Cluster size ≥ 100	1 hour fixed	1 hour sliding	3 hours fixed	3 hours sliding	6 hours fixed	6 hours sliding	12 hours fixed	12 hours sliding
Round of 16	en	RT	24,336	1,972	493	717	898	1,219	1,217	1,502	1,480	1,811
		RE	60,707	3,347	177	281	385	596	603	781	746	903
	es	RT	4,289	302	111	154	201	247	248	297	295	315
		RE	10,268	389	45	69	108	158	186	223	243	321
	pt	RT	515	18	4	9	11	16	18	21	20	21
		RE	3,153	59	1	4	7	7	8	10	10	18
	de	RT	141	22	3	4	7	11	11	13	11	13
		RE	297	16	3	3	4	5	6	5	4	6
Quarter-finals	en	RT	16,332	1,591	474	647	775	1,008	999	1,245	1,227	1,519
		RE	33,449	1,739	139	222	293	371	378	528	507	617
	es	RT	1,758	94	28	46	64	86	88	117	111	128
		RE	8,095	247	30	33	47	81	81	109	116	145
	pt	RT	299	15	5	7	7	6	5	10	9	12
		RE	4,888	98	11	15	22	18	25	31	28	31
	de	RT	120	13	3	3	4	4	6	6	8	10
		RE	116	8	0	1	1	1	1	3	3	3
Semi-finals	en	RT	18,557	1,702	446	668	841	1,095	1,115	1,425	1,424	1,703
		RE	31,597	1,443	125	180	232	311	294	404	421	504
	es	RT	4,690	421	120	177	225	283	280	343	343	392
		RE	5,845	195	25	47	68	74	76	111	97	141
	pt	RT	469	34	10	14	17	18	18	21	21	25
		RE	3,956	88	10	12	14	23	27	30	31	38
	de	RT	117	13	1	2	5	8	8	12	12	13
		RE	57	0	0	0	0	0	0	0	0	0
Final	en	RT	24,775	2,033	443	675	821	1,101	1,117	1,432	1,436	1,815
		RE	50,630	2,998	160	235	290	394	395	527	549	657
	es	RT	1,293	56	17	25	32	46	45	53	55	76
		RE	18,036	721	49	72	82	113	113	148	172	222
	pt	RT	390	26	5	8	13	16	16	23	21	22
		RE	4,197	99	4	6	12	12	14	10	18	15
	de	RT	158	7	3	4	7	9	9	12	11	14
		RE	183	2	0	1	1	1	1	3	2	7

Table 5.9: Comparison between the number of clusters with size equal or greater than 10 and the number of clusters that have ambiguous sentiment classifications in periods of sentiment homophily, for each different strategy used.

أبويافا
@abdallahsals...
Follow

While football players are running the field #ARGvsNED
Children are running for their Lives #Gaza

#GazaUnderAttack

1:47 PM - 9 Jul 2014

70 24

Number of tweets: 11, Overall sentiment: -1.81818 (from -4 to 4)

Jul-09 10:19 | 1 | -2 | 966534576: Martyr child Raneen Abdul Ghafoor, a year and three months #GAZA #GazaUnderAttack #WorldCup http://t.co/X94i3b5VIM

Jul-09 10:29 | 1 | -2 | 966534576: bombing of #Gaza by sea and land at the same time #WorldCup #GazaUnderAttack

Jul-09 10:40 | 1 | -2 | 966534576: Stop killing Gaza #GazaUnderAttack #WorldCup

Jul-09 11:34 | 1 | -4 | 966534576: Ahmed Malkii dreams of becoming a pilot .. Today Killed by Israeli Warplanes http://t.co/OPqsfdxueb #GazaUnderAttack #WorldCup

Jul-09 11:44 | 1 | -3 | 966534576: Gaza is under attack by Israeli airstrikes #WorldCup #GazaUnderAttack #GazaUnderFire http://t.co/TybXaaTIUQ

Jul-09 11:47 | 1 | -1 | 966534576: While football players are running the field #ARGvsNED Children are running for their Lives #Gaza #GazaUnderAttack http://t.co/r6nVVTIP0T

Jul-09 12:22 | 1 | -4 | 966534576: Kinan Is the only survivor of Hamad family And the rest of the family killed http://t.co/uJEiXImuXv #GazaUnderAttack #WorldCup #ARGvsNED

Jul-09 12:27 | 1 | -2 | 966534576: During the last 48 hours about :: 500 Raid on Gaza #WorldCup #GazaUnderAttack #ARGvsNED

Jul-09 12:40 | 2 | -4 | 966534576: Look at the innocent child being murdered by Israeli forces in #Gaza. http://t.co/Kunrs8SKm0 #GazaUnderAttack #ARGvsNED #WorldCup

Jul-09 13:04 | 1 | -4 | 966534576: IN #Gaza - Palestine IS THERE ANY HARDER THAN CRYING MOTHER of Martyr!! http://t.co/05ingBGGRP #GazaUnderAttack #WorldCup #ARGvsNED

Jul-09 18:49 | 1 | -4 | 966534576: All the world will wath the game #ARGvsNED and we are in #Gaza are bombed and killed our children #GazaUnderAttack #WorldCup

Figure 5.21: Example of an ambiguity, with the sentiment value of (1, -1), surrounded by a negative context.

Chapter 6

Evaluation

Our approach is divided in two different stages. The first stage comprises the integration of clustering and sentiment classification techniques, while in the second stage we propose three techniques for analyzing the extent of influence over sentiment and searching sentiment homophily in the obtained clusters.

In the first stage it is important to ascertain the quality of the resulting clusters, while in the second stage we want to validate the suggested approach for finding local sentiment homophily in clusters. The other two proposed methods are not evaluated because the method that exploits clustering in narrow periods of time revealed to be unsuitable for sentiment homophily analysis in early results, and the analysis of the polarity dynamics over time was made in an unsystematic way to guide our decisions to pursue sentiment homogeneity locally in time.

To evaluate user clustering we used the Modularity Measure, while for validating the approach followed in the Algorithms 3 and 4 we used the model K-Fold Cross Validation and Manual Validation.

6.1 Modularity Measure for User Clustering

According to Newman [70], Modularity Q is a measure about the division of the nodes in a graph into different clusters, and the strength of their connections. It tests the number of edges inside each cluster with the number of edges that would be expected if the edges were randomly distributed. The higher the value of Q the higher the density of the connections inside the cluster and the higher the sparsity among clusters.

The modularity of a configuration of c clusters from a graph with m links is defined as:

$$Q = \sum_{i=1}^{c} (e_{ii} - {a_i}^2), Q \in [-0.5, 1], \tag{6.1}$$

Where, for every pair of nodes vw, let $A_{vw} = 1$ when there is a link between v and w, $A_{vw} = 0$ otherwise, $\delta(c_v, c_w) = 1$ when v and w belong to the same cluster, $\delta(c_v, c_w) = 0$ otherwise, then we have the fraction of edges with both nodes in cluster i:

$$e_{ii} = \sum_{vw} \frac{A_{vw}}{2m} \delta(c_v, c_w), \tag{6.2}$$

And let k_i be the number of links to the cluster i, then we have the fraction of nodes connected to

Daily-based Clustering		
Cluster set	*Q (RT)*	*Q (RE)*
June 27th	0.743	0.914
June 28th	0.598	0.896
June 29th	0.635	0.901
June 30th	0.643	0.893
July 1st	0.625	0.898
July 2nd	0.725	0.917
July 3rd	0.742	0.917
July 4th	0.581	0.891
July 5th	0.605	0.850
July 6th	0.682	0.903
July 7th	0.711	0.911
July 8th	0.658	0.914
July 9th	0.623	0.893
July 10th	0.602	0.899
July 11th	0.673	0.903
July 12th	0.616	0.897
July 13th	0.597	0.895
July 14th	0.612	0.877
July 15th	0.719	0.912
Average	0.652	0.899

Round-based Clustering		
Cluster set	*Q (RT)*	*Q (RE)*
Round of 16	0.538	0.796
Quarter-finals	0.683	0.802
Semi-finals	0.607	0.824
Final	0.650	0.777
Average	0.620	0.800

Global Clustering		
Cluster set	*Q (RT)*	*Q (RE)*
Global	0.502	0.636

Table 6.1: Modularity results for each set of clusters.

other nodes in i:

$$a_i = \frac{k_i}{2m} \tag{6.3}$$

We used the modularity measure to evaluate the quality of the different set of clusters obtained by clustering the users for the entire time interval of the dataset, for each day of the knock-out stage, and for each round of the knock-out stage. With this measure we can understand the general density of the resulting clusters and how well defined these social circles are. Table 6.1 contains the respective modularity results.

The closer the modularity is to 1, more independent the clusters tend to be from each other. Lower values of Q reveal the presence of bigger clusters and that was what we found with the global clustering. The size of the larger clusters were considerably high and they represented a significant percentage of the giant component of these two graphs (retweet and reply-relations). For this reason we did not use the global clustering in the later stages.

With daily-based and round-based clustering we got better modularity results and we identified a bigger difference between the size of the larger cluster and the size of the giant component, which suggests that the giant component is divided in minor clusters.

These results may explain why we found reply-based clusters to be smaller than the retweet-based clusters, since it suggests a higher degree of division between the communities. The difference between the retweet-based and reply-based results of modularity reveal that the Layered Label Propagation can perform a better partitioning in reply-relations graphs.

6.2 Manual validation for Local Polarity Homophily

With the sentiment time-lining we may have found some moments where sentiment homophily was expressed, but we cannot state the reason nor be sure if they were not merely a result of chance. The fact that some of these moments were related with chains of reweets and conversations supported the hypothesis of influence carrying sentiment and spreading an overall sentiment polarity in clusters.

To formally identify those time intervals of sentiment homogeneity we defined the metrics time window t, minimum number of tweets m, and minimum rate of polarity prevalence p. However, this does not state whether there is an increased probability of a given user tweeting a message with the same sentiment polarity as the prevalent sentiment of the cluster to which he belongs, when t, m, and p are met. To answer this question we exploited dubious sentiment classifications, since they do not evidence a clear polarization, and we searched for those who were inside moments of apparent sentiment polarity homophily. Then, we relabeled their sentiment classification with the prevalent polarity in the cluster. Finally, we asked human coders to classify these ambiguities, and we compared the extrapolation results with the human-coders classifications.

We are using this strategy to find evidence of sentiment homophily, trying to understand whether the prevalence of a sentiment in the cluster may influence the sentiment of their individuals, increasing the chance of those individuals express a befitting sentiment during that context. Moreover, in case this hypothesis is found to be true, this same strategy can be used to improve the automatic sentiment classification, by relabeling dubious classifications according to the overall sentiment of the cluster, when it exists.

We performed two different manual classifications with human-coders. One was used to classify the sentiment of tweets which were previously classified by SentiStrength with ambiguous polarity. The other was used as control group, to which we gave tweets with non-ambiguous sentiment classifications.

For the ambiguities classification we gathered 24 human-coders, in which 23 of them are Portuguese native-speakers and the remaining one is a Spanish native-speaker. All of them are able to read and interpret English, and 18 are also able to read and interpret Spanish. We shuffled them in 8 groups of 3, and each group evaluated two sets of 100 ambiguous tweets each. This way, each ambiguity was classified by three different human-coders.

The testing samples were randomly collected from the set of ambiguous tweets found with Algorithm 4, using the fixed parameters $t = 6$, $m = 10$, and $p = 0.7$. These samples sum a total of $1,600$ ambiguous tweets, divided into 800 for English, 600 for Spanish, and 200 for Portuguese. Half of the sets of each language was extracted from retweet-based clusters, and the other half from reply-based clusters.

The samples were lined up as follows:

- Group 1: Sample of retweet-based English ambiguities, and sample of retweet-based Spanish ambiguities;
- Group 2: Sample of retweet-based English ambiguities, and sample of retweet-based Spanish ambiguities;
- Group 3: Sample of retweet-based English ambiguities, and sample of retweet-based Spanish ambiguities;
- Group 4: Sample of retweet-based English ambiguities, and sample of reply-based Spanish ambiguities;
- Group 5: Sample of reply-based English ambiguities, and sample of reply-based Spanish ambiguities;

- Group 6: Sample of reply-based English ambiguities, and sample of reply-based Spanish ambiguities;
- Group 7: Sample of reply-based English ambiguities, and sample of retweet-based Portuguese ambiguities;
- Group 8: Sample of reply-based English ambiguities, and sample of reply-based Portuguese ambiguities;

People were randomly assigned to these groups but we ensured that the Spanish-speakers were assigned to some group with a Spanish sample.

Each person was asked to classify the sentiment expressed in the tweet message, as *positive*, *neutral*, or *negative*. We chose to only ask for the polarity and not the sentiment strength to simplify the classification process. We included the *neutral* option assuming that there are indeed some tweets that do not express any kind of polarization. Alongside with the testing tweet and the classification options, we also provide the surrounding buzz of the social circle to which the tweet belongs to give contextual information that may help the coder to understand the tweet's message. However, we did not present any information about the automatic sentiment classification of any displayed tweet to not bias the human-coders' decisions. In Appendix C we show the classification environment that the coders used to perform this task.

The results of the manual evaluation for the strategy implemented with Algorithm 4 are presented in Table 6.2. At first sight we can observe the lack of unanimity among human-coders, with a rate of agreement for each group of three persons around 36%. This unveils the subjectivity nature of sentiment classification, that must be considered when we are analyzing this type of results. For this reason we analyzed the compliance of our strategy's results with the unanimous classifications of the human-coders, defining two levels of unanimity: total unanimity and agreement of at least 2 coders.

With these two types of gold standard classification results we compared them with the result of the clusters' polarity extrapolation technique, regarding the number of matches and mismatches, and also the number of neutral classifications.

These results suggest a tendency for the real sentiment of ambiguous tweets to match the overall sentiment of their clusters, over having a neutral or mismatching sentiment polarity, and this value is clearly higher than it would be assigned by chance. However, since this extrapolation matched the manual classifications just in around 50% of the cases of total unanimity among human-coders and this value was never higher than 60.42% it is not sufficient to claim that when there is a period of sentiment homophily there is a strong chance of a user in that cluster sharing a tweet with an equivalent polarity.

Regardless the language or the link-type of the clusters, the results were quite similar in what concerns the rate expressiveness of mismatches, matches and neutral classifications. The rate of matches represent half the cases, and the other half is shared between neutral occurrences and mismatches. The rate of matches revealed to be quite stable for the different evaluations performed, while the rate of mismatches and neutral classifications showed a more unstable nature, revealing considerable fluctuations between different evaluations. This strategy has a reasonable success rate when compared with a random classification, but quite low to improve significantly the automatic sentiment classification.

Our initial assumption about the sentiment labels $(1,-1),(2,-2),(3,-3),(4,-4),(5,-5)$ being ambiguous instead of neutral, seems to be considerably supported by these results, since the evaluated ambiguities were classified as neutral only in less than 25% of the times.

It is important to keep in mind some different factors that can have impact in these final results. As we mentioned before, sentiment classification is a subjective task and even for humans it is hard to find consensus. SentiStrength has itself a correctness rate not higher than 63% on average in unsupervised

Set of Ambiguities		Human-coders Agreement			Cluster Sentiment Polarity Mismatch			Neutral Sentiment			Cluster Sentiment Polarity Match		
		≥ 2	*Unanimity*	*Total Disagreement*	≥ 2	*Unanimity*	*Random*	≥ 2	*Unanimity*	*Random*	≥ 2	*Unanimity*	*Random*
en	*RT*	92.50%	36.25%	7.50%	30.00%	28.28%	29.00%	24.86%	20.69%	37.75%	45.14%	**51.03%**	33.25%
	RE	93.75%	39.00%	6.25%	16.80%	20.51%	27.00%	36.53%	33.97%	34.75%	**46.67%**	45.51%	38.25%
	Total	93.13%	37.63%	6.88%	23.36%	24.25%	28.00%	30.74%	27.57%	36.25%	45.91%	**48.17%**	35.75%
es	*RT*	86.33%	32.00%	13.67%	32.43%	32.29%	36.33%	15.83%	7.29%	30.67%	51.74%	**60.42%**	33.00%
	RE	90.33%	37.67%	9.67%	34.32%	33.63%	36.33%	17.71%	7.96%	33.00%	47.97%	**58.41%**	30.67%
	Total	88.33%	34.83%	11.67%	33.40%	33.01%	36.33%	16.79%	7.66%	31.83%	49.81%	**59.33%**	31.83%
pt	*RT*	87.00%	29.00%	13.00%	32.18%	27.59%	33.00%	29.89%	24.14%	32.00%	37.93%	**48.28%**	35.00%
	RE	95.00%	42.00%	5.00%	36.84%	42.86%	35.00%	20.00%	9.52%	37.00%	43.16%	**47.62%**	28.00%
	Total	91.00%	35.50%	9.00%	34.62%	36.62%	34.00%	24.73%	15.49%	34.50%	40.66%	**47.89%**	31.50%
Global	*RT*	89.50%	33.75%	10.50%	31.15%	29.63%	32.25%	22.21%	16.30%	34.38%	46.65%	**54.07%**	33.38%
	RE	92.63%	38.88%	7.38%	25.78%	28.30%	31.50%	27.53%	21.22%	34.38%	46.69%	**50.48%**	34.13%
	Total	91.06%	36.31%	8.94%	28.41%	28.92%	31.88%	24.91%	18.93%	34.38%	46.67%	**52.15%**	33.75%

Table 6.2: Manual evaluation results regarding the approach implemented in the Algorithm 4, and comparison with a random approach.

mode [89], and it is susceptible to particular characteristics like sarcasm. Ironic tweets may create chains of retweets or sarcastic discussions that may be wrongly classified by SentiStrength and it may lead our approach to wrongly classify an overall sentiment of a cluster.

In order to evaluate the expression of some of these factors in our work, we set another manual validation round for sentiment classification of non-ambiguous tweets, and we calculated the Krippendorff's alpha-coefficient that is a statistical measure to formally ascertain the agreement among human-coded results.

6.3 Manual validation of non-ambiguous classifications

In this second round of manual classifications we only gathered 9 human-coders that we used as a small control group to evaluate non-ambiguous sentiment classifications. The purpose of this evaluation was testing the human coders classifications for trusted SentiStrength classifications.

The coders were randomly assigned to 3 different groups, in the same line of the previous round, with the unique difference that only English tweets were used.

- Group 1: Sample of retweet-based English ambiguities, and sample of reply-based English ambiguities;

- Group 2: Sample of retweet-based English ambiguities, and sample of reply-based English ambiguities;

- Group 3: Sample of retweet-based English ambiguities, and sample of reply-based English ambiguities;

We can see the results in table 6.3 where it is noticeable a considerable increase in the matches rate. This supports our assumption about the higher confidence in a polarized sentiment classification over neutral classifications. As we have seen for cluster sentiment extrapolation, the matching results are less undeviating than the neutral and mismatch results. There seems to be an arbitrary interchangeability of expressiveness between neutral and mismatch rates among different evaluations. Considering that our approach does not distinguish neutral classifications from ambiguous ones, the observable difference in the neutral rate was expected since only polarized classifications were given to the coders, in this second validation.

Set of Polarized Tweets		Human-coders Agreement			SentiStrength Polarity Mismatch		Neutral Sentiment		SentiStrength Polarity Match	
		≥ 2	*Unanimity*	*Total Disagreement*	≥ 2	*Unanimity*	≥ 2	*Unanimity*	≥ 2	*Unanimity*
en	*RT*	93.67%	45.00%	6.33%	16.01%	14.07%	9.96%	2.96%	74.02%	82.96%
	RE	95.00%	40.00%	5.00%	11.23%	6.67%	22.81%	9.17%	65.96%	84.17%
	Total	94.33%	42.50%	5.67%	13.60%	10.59%	16.43%	5.88%	69.96%	83.53%

Table 6.3: Manual evaluation results of polarized sentiment classifications obtained with SentiStrength.

6.4 Krippendorff's alpha reliability about Human-coders Agreement

As we have seen in both manual evaluations results, there is a low agreement degree inside each group of three coders regarding the inherent sentiment of tweets. This subjective nature can bias the evaluation result itself. To evaluate the quality of the sentiment classifications made by each group of human coders we chose to calculate their Krippendorff's alpha coefficient. This measure encodes the reliability of a set of independent classifications.

Considering an evaluation set of $u_1, ..., u_N$ evaluating units and a group $h_1, ..., h_m$ of independent human-coders, where each coder h_i assigns to the unit u_j a classification value v_{ij}, we get a matrix $m * N$ of all evaluation values. When the result of all evaluations generates a complete matrix, the number n of all pairable values is $n = mN$.

The general canonical form of this measure is:

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{6.4}$$

Where D_o is the observed disagreement among the values assigned to the evaluating units, and D_e is the expected disagreement if the evaluation values were assigned by chance.

There is statistical reliability for a set of independent evaluations when

$$1 \geq \alpha \geq 0 \tag{6.5}$$

To obtain the value of the disagreement measures D_o and D_e we need to calculate a coincidence matrix of coincidence frequencies o_{ck}. Let $V = \{l_1, ..., l_w\}$ be the set of possible values of $v_{ij} \in V$, the coincidence matrix is $V * V$ and $c, k \in V$ where,

$$o_{ck} = \sum_u \frac{\gamma_u(c,k)}{m-1}, \tag{6.6}$$

Where $\gamma_u(c,k)$ is the number of (c,k) pairs in the set of all classification values $\{v_{1u}, ..., v_{mu}\}$ assigned to unit u. When $c = k$, it is important to recall that these are ordered pairs and each two matching values will originate two (c,k) pairs.

According to Krippendorff [54], this measure is a generalization of several reliability indices, being deductible into different forms depending on the number of coders, the completeness of the evaluation data, the number, scale and level of measure of the classification categories.

Let n_c be the frequency of coincidences for $c \in V$,

$$n_c = \sum_k o_{ck} \tag{6.7}$$

The proper variation of α for our set of nominal classifications (negativity, neutrality, positivity) of

chunks of 100 tweets performed by 3 coders is defined as follows:

$$\alpha_{nominal} = 1 - \frac{D_o}{D_e} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \tag{6.8}$$

The $\alpha - reliability$ of each evaluated set resulting from the manual validation of ambiguous sentiment classifications is:

- Group 1: $\alpha_{en,rt} = 0.40099$, $\alpha_{es,rt} = 0.36961$
- Group 2: $\alpha_{en,rt} = 0.38857$, $\alpha_{es,rt} = 0.26990$
- Group 3: $\alpha_{en,rt} = 0.25219$, $\alpha_{es,rt} = 0.30920$
- Group 4: $\alpha_{en,rt} = 0.45571$, $\alpha_{es,re} = 0.41684$
- Group 5: $\alpha_{en,re} = 0.53167$, $\alpha_{es,re} = 0.44423$
- Group 6: $\alpha_{en,re} = 0.30900$, $\alpha_{es,re} = 0.24703$
- Group 7: $\alpha_{en,re} = 0.32671$, $\alpha_{pt,rt} = 0.33109$
- Group 8: $\alpha_{en,re} = 0.38950$, $\alpha_{pt,re} = 0.42334$

And the $\alpha - reliability$ of each evaluated set resulting from the manual validation of non-ambiguous sentiment classifications is:

- Group 1: $\alpha_{en,rt} = 0.23300$, $\alpha_{en,mt} = 0.20947$
- Group 2: $\alpha_{en,rt} = 0.38485$, $\alpha_{en,mt} = 0.39672$
- Group 3: $\alpha_{en,rt} = 0.40963$, $\alpha_{en,mt} = 0.10551$

These results show that the coders evaluation is statistically reliable, but far from perfect reliability, with a considerable level of disagreement. This may indicate that, given the subjectivity of this task, it would be desirable to have a higher odd number of human-coders per evaluation set.

6.5 K-fold Cross Validation

Our strategy of extrapolating the overall sentiment of clusters to individual ambiguous classifications works like a predictive model that uses the sentiment prevalence for estimation. Although Algorithms 3 and 4 look for moments of sentiment prevalence, they may include a certain rate of ambiguities and opposite polarities. To assess the extent of sentiment prevalence when it exists we used K-fold Cross Validation technique over those periods of time.

According to Webb and Copsey [106], for estimating the error rate E of the model given, a set of m testing examples, a value for the parameter k it is initially defined. Then, the testing examples are divided into k folds, which are chunks of approximately m/k testing examples arranged in a random order. For k iterations, a different fold i is used for testing and the remaining $k-1$ folds are used for estimating the classification of i, in which the number n_i of wrongly classified examples is measured. The error rate is given by:

$$E = \frac{\sum_{i=1}^{k} n_i}{m} \tag{6.9}$$

Stage	*Language*	*Type*	*Fixed time window*				*Sliding time window*			
			$t=1$	$t=3$	$t=6$	$t=12$	$t=1$	$t=3$	$t=6$	$t=12$
Round of 16	*en*	*RT*	15.47%	16.64%	17.46%	17.62%	16.60%	17.64%	17.99%	18.41%
		RE	20.55%	22.04%	22.38%	22.73%	22.55%	22.61%	23.06%	22.92%
	es	*RT*	14.63%	17.93%	18.78%	18.61%	17.82%	18.97%	19.23%	19.01%
		RE	21.55%	20.87%	20.57%	19.47%	21.18%	20.04%	19.80%	18.55%
	pt	*RT*	10.00%	13.36%	13.90%	16.51%	14.20%	15.71%	17.32%	17.17%
		RE	30.00%	25.00%	24.00%	21.39%	22.50%	22.50%	24.17%	23.82%
Quarter-finals	*en*	*RT*	14.98%	15.23%	15.48%	16.52%	15.68%	16.08%	17.11%	17.67%
		RE	19.69%	20.33%	20.18%	21.76%	20.36%	21.34%	22.65%	22.20%
	es	*RT*	13.81%	14.97%	14.34%	15.72%	13.42%	15.35%	17.14%	17.65%
		RE	22.24%	21.59%	22.48%	22.93%	21.64%	22.98%	22.68%	23.62%
	pt	*RT*	14.88%	14.69%	11.97%	15.20%	17.59%	14.01%	15.78%	15.45%
		RE	18.75%	22.74%	22.35%	21.38%	20.38%	20.95%	22.09%	21.36%
Semi-finals	*en*	*RT*	14.75%	16.23%	16.56%	17.12%	16.35%	16.98%	17.40%	17.83%
		RE	19.54%	19.55%	20.13%	20.62%	20.50%	20.82%	21.34%	21.30%
	es	*RT*	15.15%	17.15%	16.64%	17.67%	16.82%	18.06%	18.40%	18.66%
		RE	20.68%	23.27%	21.98%	21.80%	22.84%	23.28%	23.29%	23.06%
	pt	*RT*	16.83%	14.61%	15.85%	16.70%	17.14%	14.94%	17.57%	18.31%
		RE	18.13%	16.61%	22.06%	22.80%	16.95%	21.88%	22.84%	24.62%
Final	*en*	*RT*	13.78%	14.48%	15.00%	16.09%	14.81%	15.50%	16.44%	17.57%
		RE	17.72%	19.91%	20.06%	21.22%	20.04%	21.28%	21.48%	21.87%
	es	*RT*	19.09%	14.14%	16.96%	16.10%	16.75%	18.17%	17.60%	18.24%
		RE	22.79%	22.67%	23.69%	23.73%	24.22%	22.47%	23.76%	23.83%
	pt	*RT*	11.03%	15.42%	14.10%	15.21%	18.25%	15.77%	15.24%	15.76%
		RE	18.38%	22.78%	25.75%	25.69%	18.10%	25.98%	26.26%	24.63%
%			**17.69%**	**18.43%**	**18.86%**	**19.36%**	**18.61%**	**19.30%**	**20.03%**	**20.15%**

Table 6.4: Error rate E average of K-Fold Cross Validation, for $k = 10$, over sets of tweets in periods of prevalence of a certain sentiment polarity.

We adapted Algorithms 3 and 4 to return the set of all tweets in the cluster tweeted in t alongside with the ambiguity. This way, we obtained the entire context of the cluster in periods of prevalence of a certain sentiment polarity. We used each set of these tweets as testing examples for different runs of K-fold Cross Validation. The results were averaged regarding the testing configuration. We tested both algorithms, and $t = 1, t = 3, t = 6, t = 12$ for each retweets and replies of English, Spanish and Portuguese, during round of 16, quarter-finals, semi-finals and final. We did not considered German clusters because they show few periods of sentiment prevalence.

We set $k = 10$, except when $m < k$ where we defined $k = m$. Considering the results in Table 6.4, we can see that as long as t increases the error also increases, meaning that the extent of sentiment prevalence decreases. The error rate also seems to be slightly better when Algorithm 3 is used. However, the most noticeable difference is found between retweets and replies. This could be related with the phenomenon of retweet' chains because it generates large amounts of similar messages with similar sentiment values, which may originate more sentiment homogeneity. These findings seem to be quite uniform among the different stages and languages.

Chapter 7

Conclusion

In this document we present (i) the integration of several existing techniques for extracting social interactions from Twitter, building and clustering the network's graph from the relations on those interactions; (ii) classifying the inherent sentiment of their tweets' messages; (iii) and then we propose and evaluate three different approaches to study the overall sentiment of social circles over time, focusing on Influence and Homophily patterns.

The motivation for this analysis was first understand whether social groups tend to exhibit a prevalent sentiment and how frequently different sentiment polarities appear on those clusters. Then, we aimed to look for sentiment changes and patterns of sentiment dynamics over time. Finally, we searched for evidence of sentiment contagion to perceive whether the existence of sentiment homophily on a cluster may influence the sentiment of their individuals.

Like Nichols et al. [71] and Gruzd et al. [38] argued, we found on a big sporting event a prolific source of social data on Twitter, in our case the 2014 FIFA World Cup. Our dataset of $339,702,345$ tweets contained a suitable amount of interactions in different languages where we found clusters of English, Spanish, Portuguese and German-speaking users.

These clusters were obtained from graphs based on two different types of social interactions, characteristic from Twitter: retweets and replies. Both are considered strong ties, but while retweets appear to be more content-related, assuming a major role on information propagation, replies show direct conversations between users. For higher values of modularity we found clusters with smaller size, what confirmed that replies tend to be more restrict than retweets. The advantage of using replies instead of all mentions was allowing an independent analysis between retweets and replies, because they are mutually exclusive.

WebGraph and LAW frameworks were suitable to deal with large networks' graphs, and the Layered Label Propagation algorithm performed well when clustering graphs from sets of relations with time windows from 1 to 6 days, showing good modularity results. However, when clustering the complete graph of the entire dataset, it returned larger clusters with lower modularity coefficient. The distribution of the obtained clusters by their size followed a power-law, either for clusters of users or their respective clusters of tweets.

SentiStrength also offered an efficient solution to classify large amounts of tweets. Its versatility allowed us to classify tweets in English, Spanish, Portuguese and German, by using different configuration files provided by the SentiStrength's author. Spanish, Portuguese and German files were adaptations made by students and we made some improvements on the Portuguese files. The major drawback of using SentiStrength was the fact that it classifies as neutral $(1, -1)$ both neutral and undecidable sentences, without distinction.

Regarding our first hypothesis about the existence of sentiment prevalence on clusters, we started

by testing whether using a time window of 24 hours for clustering would be enough to detect sentiment trends. However, we found that the majority of the clusters denote sentiment variations during the day and their distribution of sentiment values reveals sentiment heterogeneity, even for polarity, like Thelwall et al. [88] also found. Nevertheless, the most frequent sentiment classification was $(1, -1)$, followed by $(2, -1)$, independently of the day, language or relation type in the origin of the clusters.

For analyzing the sentiment dynamics over time we disregarded the daily clustering and we used clusters obtained with the relations of each round of the World Cup's knock-out stage. This way, we covered a broader range of time, maintaining good clustering results. For this analysis we time-lined the frequency of absolute sentiment values (the sum of the positive with the negative value obtained with SentiStrength) for each hour, during each cluster's time-life. Giant clusters revealed coincident bursts of sentiment among the different sentiment values, probably because these clusters could be pruned in smaller clusters, while very small clusters with sparse tweets in time gave few sentiment information for reasoning about sentiment homophily. Nonetheless, the majority of clusters have sentiment spikes during their time-life and we detected that chains of polarized retweets generate moments of sentiment homogeneity, as well as some topic-related conversations, which are respectively more frequent (but not exclusively) on retweet-based clusters, and reply-based clusters. If we assume that, when some user retweets a certain status, there is a chance of that user being also sharing the inherent sentiment of that status' message, then we may say that there is sentiment influence on cascades of retweets. These spikes of sentiment polarity are volatile on time and they usually last only a few hours. Periods of sentiment homophily on retweet-based clusters seem to have bigger spikes and a quieter surrounding, while reply-based clusters show less detached spikes with a noisier surrounding.

By time-lining the absolute sentiment frequencies we informally observed that periods of polarized sentiment occur in ranges of a few hours. To formally detect these moments of sentiment homogeneity we defined a metric that includes a time window t, a minimum number of tweets m during t, and a minimum rate p of a prevalent sentiment polarity during t. However, this metric alone does not give insight about whether a certain user inside that cluster is more susceptible to share the same sentiment polarity as his cluster when these metrics are met, or not.

Therefore, we exploited ambiguous sentiment classifications for testing if the prevalent sentiment of clusters can be extrapolated for their individuals. We implemented two similar strategies for finding ambiguities in those conditions using t, m, p, which differ in the positioning of the ambiguity in t. One strategy searches for ambiguities with a fixed position in the middle of the time window, while the other looks for ambiguities in any time in t. Those ambiguities were relabeled with the prevalent sentiment in the cluster in that time and then they were manually classified by a group of human-coders.

The result of the manual validation demonstrated a certain level of disagreement among the coders, but statistically reliable according to the Krippendorff's alpha coefficient. The matching rate between the human-coders classification and the clusters' sentiment polarity extrapolation always showed the higher and more stable expressiveness over mismatching and neutral rates. However, with the best matching result around 60%, we can say that, as Thelwall [85] found weak but significant level of sentiment homophily among direct following users, we also found a weak but statistically significant tendency of an user sharing a befitting sentiment in a cluster during a period of sentiment homogeneity. The K-fold Cross Validation unveiled that this homogeneity is usually stronger in retweet-based clusters.

Chapter 8

Future Work

Our work explores the combination of different techniques for reasoning about sentiment patterns and dynamic behaviors at a cluster level and their relation with Influence and the possibility of Sentiment Homophily.

We identified some drawbacks of certain decisions that may have bias the final results. For instance, the inner accuracy of SentiStrength on unsupervised mode is around 60% [89] which should be taken into account during the evaluation of strategies that include this tool. We have considered SentiStrength neutral classifications as ambiguous classifications because it classifies undecidable sentences as neutral, as well as it classifies sentences which it clearly identifies as having no polarity strength according to its lexicon. It would be valuable to clearly identify when SentiStrength does not recognize any word in the sentence with its lexicon. This way we would only consider undecidable tweets and the ambiguities $(2, -2), (3, -3), (4, -4), (5, -5)$ for the clusters' sentiment polarity extrapolation.

Given the subjective nature of sentiment classification, it would be desirable to have an higher odd number of coders for each set in the manual validation.

Regarding Algorithm 3, it would be interesting to test a variant where the ambiguity appears in the end of the time window, and another where it appears in the beginning. This would be useful for reasoning about Sentiment Influence, since in the beginning of the time window the exposure to that sentiment is expected to be lower than at the end.

With our approach we extracted the information about several clusters of different sizes and characteristics. Inside these clusters we have searched for moments of sentiment homophily and we extrapolated that information for dubious cases inside that context. Another interesting approach would be to find arbitrary tweets with dubious classifications in the dataset, building the ego-networks of their users using local clustering techniques, investigating the sentiment in these networks during a certain time window that includes the moment when those statuses were tweeted and ascertain how frequently a prevalent sentiment is found and how successful would be assuming the overall sentiment of the clusters surrounding those ambiguities.

Bibliography

[1] O. Alonso, C. C. Marshall, and M. Najork. Are some tweets more interesting than others? #hardquestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 2:1–2:10, New York, NY, USA, 2013. ACM.

[2] L. R. Anderson and C. A. Holt. Information cascades in the laboratory. *American Economic Review*, 87:847–862, 1995.

[3] I. Anger and C. Kittl. Measuring influence on twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, pages 31:1–31:4, New York, NY, USA, 2011. ACM.

[4] T. Antal, P. L. Krapivsky, and S. Redner. Social balance on networks: The dynamics of friendship and enmity. *Physica D: Nonlinear Phenomena*, 224(1-2):130–136, Dec 2006.

[5] M. Arias, A. Arratia, and R. Xuriguera. Forecasting with twitter data. *ACM Trans. Intell. Syst. Technol.*, 5(1):8:1–8:24, Jan 2014.

[6] A. Asiaee T., M. Tepper, A. Banerjee, and G. Sapiro. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1602–1606, New York, NY, USA, 2012. ACM.

[7] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[8] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[9] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[10] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 519–528, New York, NY, USA, 2012. ACM.

[11] P. S. Bearman and J. Moody. Suicide and friendships among american adolescents. *American journal of public health*, 94(1):89–95, Jan 2004.

[12] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09, pages 231–236, Washington, DC, USA, 2009. IEEE Computer Society.

[13] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 587–596, New York, NY, USA, 2011. ACM.

[14] P. Boldi and S. Vigna. The webgraph framework i: Compression techniques. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 595–602, New York, NY, USA, 2004. ACM.

[15] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artif. Life*, 17(3):237–251, Aug 2011.

[16] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[17] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu. What is tumblr: A statistical overview and comparison. *SIGKDD Exploration Newsletter*, 16(1):21–29, Sept. 2014.

[18] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.

[19] C. Chew and G. Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118, 11 2010.

[20] M. M. F. Chowdhury, M. Guerini, S. Tonelli, and A. Lavelli. Fbk: Sentiment analysis in twitter with tweetsted. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 466–470. Association for Computational Linguistics, 2013.

[21] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA, 2010. ACM.

[22] R. Cuevas, R. Gonzalez, A. Cuevas, and C. Guerrero. Understanding the locality effect in twitter: Measurement and analysis. *Personal Ubiquitous Comput.*, 18(2):397–411, Feb 2014.

[23] A. O. Durahim and M. Coşkun. #iamhappybecause: Gross national happiness through twitter analysis and big data. *Technological Forecasting and Social Change*, 99:92 – 105, 2015.

[24] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[25] I. Eleta. Multilingual use of twitter: Social networks and language choice. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 363–366, New York, NY, USA, 2012. ACM.

[26] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422, 2006.

[27] Facebook. Facebook - company info, 2015. [Online at `http://newsroom.fb.com/company-info/`; accessed 2015-September-19].

[28] Facebook. Facebook - localization and translation, 2015. [Online at `http://newsroom.fb.com/products/`; accessed 2015-September-19].

[29] Facebook. Facebook - newsroom, 2015. [Online at `http://newsroom.fb.com/news/2014/07/world-cup-breaks-facebook-records/`; accessed 2015-September-19].

[30] R. Fan, J. Zhao, Y. Chen, and K. Xu. Anger is more influential than joy: Sentiment correlation in weibo. *PLoS ONE*, 9(10):e110184, 10 2014.

[31] J. Fowler and N. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *British Medical Journal*, 337:a2338, 2008.

[32] A. Go, R. Bhayani, and L. Huang. Sentiment140: Stanford twitter sentiment. `http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip`, Jun 2009.

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

[34] Google+. Google+ - activity, 2015. [Online at `http://googleblog.blogspot.pt/2013/10/google-hangouts-and-photos-save-some.html`; accessed 2015-September-19].

[35] Google+. Google+ - business, 2015. [Online at `https://www.google.com/work/apps/business/products/googleplus/`; accessed 2015-September-19].

[36] Google+. Google+ - languages, 2015. [Online at `https://support.google.com/plus/answer/1044955?hl=en`; accessed 2015-September-19].

[37] O. Goonetilleke, T. Sellis, X. Zhang, and S. Sathe. Twitter analytics: A big data management perspective. *SIGKDD Explorations Newsletter*, 16(1):11–20, Sept. 2014.

[38] A. Gruzd, S. Doiron, and P. Mai. Is happiness contagious online? a case of twitter and the 2010 winter olympics. In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences*, HICSS '11, pages 1–9, Washington, DC, USA, 2011. IEEE Computer Society.

[39] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.

[40] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[41] U. R. Hodeghatta. Sentiment analysis of hollywood movies on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1401–1404, New York, NY, USA, 2013. ACM.

[42] C. Honey and S. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–10, Jan 2009.

[43] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 537–546, New York, NY, USA, 2013. ACM.

[44] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[45] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2008.

[46] Instagram. Instagram - homepage, 2015. [Online at `https://www.instagram.com`; accessed 2015-September-19].

[47] Instagram. Instagram - press, 2015. [Online at `https://instagram.com/press/`; accessed 2015-September-19].

[48] J. Ito, T. Hoshide, H. Toda, T. Uchiyama, and K. Nishida. What is he/she like?: Estimating twitter user attributes from contents and social neighbors. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1448–1450, New York, NY, USA, 2013. ACM.

[49] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, Nov. 2009.

[50] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[51] A. Joshi, A. R. Balamurali, P. Bhattacharyya, and R. Mohanty. C-feel-it: A sentiment analyzer for micro-blogs. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, HLT '11, pages 127–132, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[52] A. Jungherr. Tweets and votes, a special relationship: The 2009 federal election in germany. In *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, PLEAD '13, pages 5–14, New York, NY, USA, 2013. ACM.

[53] P. Kostkova, M. Szomszor, and C. St. Louis. #swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. *ACM Trans. Manage. Inf. Syst.*, 5(2):8:1–8:25, Jul 2014.

[54] K. Krippendorff. Computing krippendorff's alpha reliability. Technical report, University of Pennsylvania, Annenberg School for Communication, Jun 2011.

[55] P. Lai. Extracting strong sentiment trends from twitter. 2010.

[56] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM.

[57] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM.

[58] J. Li, X. Wang, and E. Hovy. What a nasty day: Exploring mood-weather relationship from twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1309–1318, New York, NY, USA, 2014. ACM.

[59] K. H. Lim and A. Datta. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd International Workshop on Modeling Social Media*, MSM '12, pages 25–32, New York, NY, USA, 2012. ACM.

[60] Y.-R. Lin and D. Margolin. The ripple of fear, sympathy and solidarity during the boston bombings. *EPJ Data Science*, 3(1):31, 2014.

[61] LinkedIn. Linkedin - about, 2015. [Online at `https://press.linkedin.com/about-linkedin`; accessed 2015-September-19].

[62] LinkedIn. Linkedin - activity, 2015. [Online at `http://blog.linkedin.com/2015/07/09/1-million-linkedin-publishers/`; accessed 2015-September-19].

[63] LinkedIn. Linkedin - mission, 2015. [Online at `https://www.linkedin.com/about-us`; accessed 2015-September-19].

[64] Z. Ma, A. Sun, Q. Yuan, and G. Cong. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 999–1008, New York, NY, USA, 2014. ACM.

[65] C. Marlow. Overstated - maintained relationships on facebook, 2015. [Online at `http://overstated.net/2009/03/09/maintained-relationships-on-facebook`; accessed 2015-September-19].

[66] J. Mcauley and J. Leskovec. Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data*, 8(1):4:1–4:28, Feb 2014.

[67] S. Milgram, L. Bickman, and L. Berkowitz. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, 13:79–82, 1969.

[68] T. Miyagawa. Anyevent-twitter-stream-0.27, 2015. [Online at `http://search.cpan.org/~miyagawa/AnyEvent-Twitter-Stream-0.27/`; accessed 2015-September-19].

[69] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 913–924, New York, NY, USA, 2014. ACM.

[70] M. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[71] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA, 2012. ACM.

[72] Pinterest. Pinterest - activity, 2015. [Online at https://blog.pinterest.com/en-gb/top-10-uk-pins-0; accessed 2015-September-19].

[73] Pinterest. Pinterest - homepage, 2015. [Online at https://www.pinterest.com; accessed 2015-September-19].

[74] Pinterest. Pinterest - press, 2015. [Online at https://about.pinterest.com/pt-pt/press/press; accessed 2015-September-19].

[75] Pinterest. Pinterest - visitors, 2015. [Online at https://blog.pinterest.com/pt-br/j%C3%A1-pensou-em-estagiar-no-pinterest; accessed 2015-September-19].

[76] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[77] X. Shuai, X. Liu, T. Xia, Y. Wu, and C. Guo. Comparing the pulses of categorical hot events in twitter and weibo. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 126–135, New York, NY, USA, 2014. ACM.

[78] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: Growth and dynamics in twitter follower markets. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 163–176, New York, NY, USA, 2013. ACM.

[79] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[80] J. Tang, Y. Chang, and H. Liu. Mining social media with social theories: A survey. *SIGKDD Explorations Newsletter*, 15(2):20–29, June 2014.

[81] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 53–62, New York, NY, USA, 2013. ACM.

[82] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 743–752, New York, NY, USA, 2012. ACM.

[83] M. Thelwall. *Heart and soul: Sentiment strength detection in the social web with SentiStrength (summary book chapter) (in press)*.

[84] M. Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

[85] M. Thelwall. Emotion homophily in social network site messages. *First Monday*, 15(4), 2010.

[86] M. Thelwall. Sentistrength - performance, 2015. [Online at http://sentistrength.wlv.ac.uk/performance.html; accessed 2015-September-19].

[87] M. Thelwall and K. Buckley. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617, 2013.

[88] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418, Feb 2011.

[89] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[90] V. Traag and J. Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.

[91] J. Travers, S. Milgram, J. Travers, and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[92] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. 2010.

[93] Tumblr. Tumblr - about, 2015. [Online at `https://www.tumblr.com/about`; accessed 2015-September-19].

[94] Tumblr. Tumblr - press, 2015. [Online at `https://www.tumblr.com/press`; accessed 2015-September-19].

[95] Twitter. About twitter - twitter.com, 2015. [Online at `https://about.twitter.com/company`; accessed 2015-September-19].

[96] Twitter. Twitter blog - 2014 year on twitter, 2015. [Online at `https://blog.twitter.com/2014/the-2014-yearontwitter`; accessed 2015-September-19].

[97] Twitter. Twitter blog - insights into the worldcup conversation on twitter, 2015. [Online at `https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter`; accessed 2015-September-19].

[98] Twitter. Twitter developer - post statuses/filter, 2015. [Online at `https://dev.twitter.com/streaming/reference/post/statuses/filter`; accessed 2015-September-19].

[99] Twitter. Twitter developer - rest api, 2015. [Online at `https://dev.twitter.com/rest/public`; accessed 2015-September-19].

[100] Twitter. Twitter developer - streaming api, 2015. [Online at `https://dev.twitter.com/streaming/overview`; accessed 2015-September-19].

[101] Twitter. Twitter support - posting links in a tweet, 2015. [Online at `https://support.twitter.com/articles/78124#`; accessed 2015-September-19].

[102] Twitter. Twitter support - using hashtags on twitter, 2015. [Online at `https://support.twitter.com/articles/49309#`; accessed 2015-September-19].

[103] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, New York, NY, USA, 2011. ACM.

[104] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[105] M. Watanabe and T. Suzumura. How social network is evolving?: A preliminary study on billion-scale twitter network. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 531–534, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[106] A. R. Webb and K. D. Copsey. *Statistical pattern recognition*, chapter 13.1.2. Wiley and Sons Publishing, 3rd edition, 2011.

[107] I. Weber, A. Ukkonen, and A. Gionis. Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 613–622, New York, NY, USA, 2012. ACM.

[108] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[109] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[110] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.

[111] M. Ye, X. Liu, and W.-C. Lee. Exploring social influence for recommendation: A generative model approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 671–680, New York, NY, USA, 2012. ACM.

[112] J. Yin, S. Karimi, B. Robinson, and M. Cameron. Esa: Emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2701–2703, New York, NY, USA, 2012. ACM.

[113] H. Zhang, N. Parikh, G. Singh, and N. Sundaresan. Chelsea won, and you bought a t-shirt: Characterizing the interplay between twitter and e-commerce. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 829–836, New York, NY, USA, 2013. ACM.

Appendix A

Twitter Data Keywords

Keywords related to the 2014 FIFA World Cup, used for topic-based data extraction using Twitter Streaming API:

Copa do Mundo, Campeonato do Mundo, World Cup, Coupe du Monde, Fussball-Weltmeisterschaft, Coupe du monde, Coppa del Mondo, Copa2014, Brasil2014, WorldCup, WorldCup2014, WorldCupBrazil2014, ModoBrasil2014, WC2014, Futebol, Fútbol, Fussball, Football, Soccer, Futbol, Socceroos, goal, FIFA, FIFAWORLDCUP, FiFa2014, FM2014, Joachim Löw, Halilhodzic, Alejandro Sabella, Postecoglou, Georges Leekens, Safet Susic, Scolari, Volker Finke, Sampaoli, Pekerman, Myungbo, Lamouchi, Jorge Luis Pinto, Niko Kovac, Reinaldo Rueda, del Bosque, Klinsmann, Didier Deschamps, Kwesi Appiah, Fernando Santos, van Gaal, Fernando Suárez, Roy Hodgson, Carlos Queiroz, Prandelli, Zaccheroni, Miguel Herrera, Stephen Keshia, Paulo Bento, Fabio Capello, Ottmar Hitzfeld, Oscar Tabárez, #Selecção, #LaSelección, #laseleccion, #Selecao, #selecçãoportuguesa, Neuer, Lahm, Schweinsteiger, Mesut Ozil, Thomas Müller, #Müller, Marco Reus, Toni Kroos, Mario Gotze, #Gotze, Bougherra, Feghouli, Medhi Lacen, Messi, Sergio Agüero, #Agüero, Higuaín, Mascherano, Di Maria, #Di #Maria, Tim Cahill, Lucas Neill, Brett Holman, Nacer Chadli, Eden Hazard, #Hazard, Lukaku, de Bruyne, #Bruyne, Thibaut Courtois, Vincent Kompany, Thomas Vemaelen, Witsel, Fellaini, Defour, Dzeko, Ibisevic, Pjanic, Begovic, Neymar, Dani Alves, Thiago Silva, David Luiz, Samuel Eto'o, #Eto'o, N'Koulou, Ekotto, Chedjou, Makoun, Stephane Mbia, Alexis Sánchez, Eduardo Vargas, Matías Fernández, Arturo Vidal, #Vidal, Gary Medel, Valdivia, Marcelo Díaz, Beausejour, Radamel Falcao, Mario Yepes, James Rodríguez, Teo Gutiérrez, Drogba, Salomon Kalou, Zokora, Yaya Touré, Gervinho, Eboué, Kolo Touré, Bryan Ruiz, Saborío, Bolaños, Keylor Navas, Darijo Srna, Modric, Kranjcar, Eduardo Silva, Ivica Olic, Mladen Petric, Mandzukic, Antonio Valencia, Cristian Noboa, Caicedo, Jefferson Montero, Edison Méndez, Walter Avoyí, Segundo Castillo, Iniesta, Xabi Alonso, Iker Casillas, #Casillas, Sergio Ramos, Gerard Piqué, Jordi Alba, Landon Donovan, Onyewu, Michael Bradley, Altidore, Tim Howard, Clint Dempsey, Lloris, Mandanda, Eric Abidal, Koscielny, Raphaël Varane, #Varane, Ribéry, Benzema, Olivier Giroud, #Giroud, Michael Essien, Sulley Muntari, Andre Ayew, Asamoah, Boateng, Asamoah Gyan, Karagounis, Salpingidis, Mitrouglou, Theofanis Gekas, Samaras, #Robben, #sneijder, Van der Vaart, Huntelaar, Affelay, Dirk Kuyt, van der Wiel, Heitinga, Izaguirre, Noel Valladares, Wilson Palacios, Carlo Costly, Jerry Bengtson, #bengtson, Wayne Rooney, #Rooney, Steven Gerrard, #Gerrard, Frank Lampard, #Lampard, Nekounam, Masoud Shojaei, Dejagah, Teymourian, Nosrati, Buffon, Pirlo, Ranocchia, Verratti, Shaarawy, Giuseppe Rossi, Balotelli, Pablo Osvaldo, Keisuke Honda, Kagawa, Okazaki, Yasuhito Endo, Atsuto Uchida, Javier Hernández, Chicharito, Giovani dos Santos, Andrés Guardado, Héctor Moreno, Keshi, John Obi Mikel, Vincent Enyeama, Victor Moses, Ahmed Musa, Emenike, Cristiano Ronaldo, #CR7, #Ronaldo, #VivaRonaldo, #Pepe, #BrunoAlves, Coentrão, Moutinho, #FORÇAPORTUGAL, Akinfeev, Ignashevich, Shirokov, Faizulin, kerzhakov, Benaglio, Bar-

netta, Gökhan Inler, Philippe Senderos, Xherdan Shaqiri, Fabian Schär, Granit Xhaka, Valentin Stocker, Luis Suárez, Edinson Cavani, Diego Forlán, Diego Lugano, Blatter, Platini.

Appendix B

Structure of JSON encoded Tweets

Example of a simple tweet:

```
{
  "created_at": "Thu Jun 19 21:30:34 +0000 2014",
  "id": 479738033672306700,
  "id_str": "479738033672306688",
  "text": "If we beat Costa Rica i want a kiss,
   obviously on the cheek, from the UK Queen.",
  ...
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 1432977446,
    "id_str": "1432977446",
    "name": "Mario Balotelli",
    "screen_name": "FinallyMario",
    ...
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 172257,
  "favorite_count": 94559,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": []
  },
  "favorited": false,
```

```
  "retweeted": false,
  "lang": "en"
}
```

Example of a retweet of the previous tweet:

```
{
  "created_at": "Wed Oct 14 15:43:04 +0000 2015",
  "id": 654321533414936600,
  "id_str": "654321533414936576",
  "text": "RT @FinallyMario: If we beat Costa Rica i want a kiss,
   obviously on the cheek, from the UK Queen.",
  ...
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 3291104166,
    "id_str": "3291104166",
    "name": "edson ",
    "screen_name": "edsonnebra",
    ...
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "retweeted_status": {
    "created_at": "Thu Jun 19 21:30:34 +0000 2014",
    "id": 479738033672306700,
    "id_str": "479738033672306688",
    "text": "If we beat Costa Rica i want a kiss,
     obviously on the cheek, from the UK Queen.",
    ...
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
      "id": 1432977446,
      "id_str": "1432977446",
      "name": "Mario Balotelli",
      "screen_name": "FinallyMario",
      ...
    },
```

```
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "is_quote_status": false,
    "retweet_count": 172258,
    "favorite_count": 94560,
    "entities": {
      "hashtags": [],
      "symbols": [],
      "user_mentions": [],
      "urls": []
    },
    "favorited": false,
    "retweeted": false,
    "lang": "en"
  },
  "is_quote_status": false,
  "retweet_count": 172258,
  "favorite_count": 0,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [
      {
        "screen_name": "FinallyMario",
        "name": "Mario Balotelli",
        "id": 1432977446,
        "id_str": "1432977446",
        "indices": [
          3,
          16
        ]
      }
    ],
    "urls": []
  },
  "favorited": false,
  "retweeted": false,
  "lang": "en"
}
```


Example of a reply to the first tweet:


```
{
  "created_at": "Thu Jun 19 21:44:31 +0000 2014",
  "id": 479741543973408800,
  "id_str": "479741543973408768",
```

```
  "text": "@FinallyMario: If we beat Costa Rica i want a kiss,
   obviously on the cheek, from the UK Queen. – Wouldn't even
   surprise me..",
  ...
  "in_reply_to_status_id": 479738033672306700,
  "in_reply_to_status_id_str": "479738033672306688",
  "in_reply_to_user_id": 1432977446,
  "in_reply_to_user_id_str": "1432977446",
  "in_reply_to_screen_name": "FinallyMario",
  "user": {
    "id": 436702610,
    "id_str": "436702610",
    "name": "Vincent Kompany",
    "screen_name": "VincentKompany",
    ...
  },
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "is_quote_status": false,
  "retweet_count": 3463,
  "favorite_count": 2707,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [
      {
        "screen_name": "FinallyMario",
        "name": "Mario Balotelli",
        "id": 1432977446,
        "id_str": "1432977446",
        "indices": [
          1,
          14
        ]
      }
    ],
    "urls": []
  },
  "favorited": false,
  "retweeted": false,
  "lang": "en"
}
```

Appendix C

Sentiment Polarity Classification for Human-coders

Task description as it was asked to the human-coders to perform the manual sentiment polarity classification:

"Olá, muito obrigado por aceitares este desafio! Vou-te pedir para classificares o sentimento que consideras estar inerente a cada tweet que te vou mostrar. Todos eles foram tweetados durante o Mundial de Futebol de 2014 e pretendo perceber se eles expressam algo positivo, negativo ou se não demonstram nenhum sentimento evidente. À esquerda vão surgir o tweet por classificar e as três classificações possíveis do seu sentimento: negativo, neutro, positivo. Assim que escolheres uma opção, surgirá um novo por classificar, até concluíres. Se o tweet ainda estiver disponível na web ele surgirá completo na janela, caso contrário aparecerá somente o seu texto. À direita vão aparecer alguns tweets que foram tweetados por utilizadores do mesmo círculo social da pessoa do tweet à esquerda, que poderão dar algum contexto ao tweet a avaliar (que estará destacado com sublinhado cinzento). O tempo que demoras não interessa para o estudo, mas 15 minutos deverão ser suficientes para terminares. Diverte-te!"

Figures C.1 and C.2 show the environment where the human-coders classified the sentiment polarity of a given set of tweets.

Figure C.1: Classification environment for human-coders evaluate the sentiment polarity of the displayed tweet as positive, neutral, or negative.

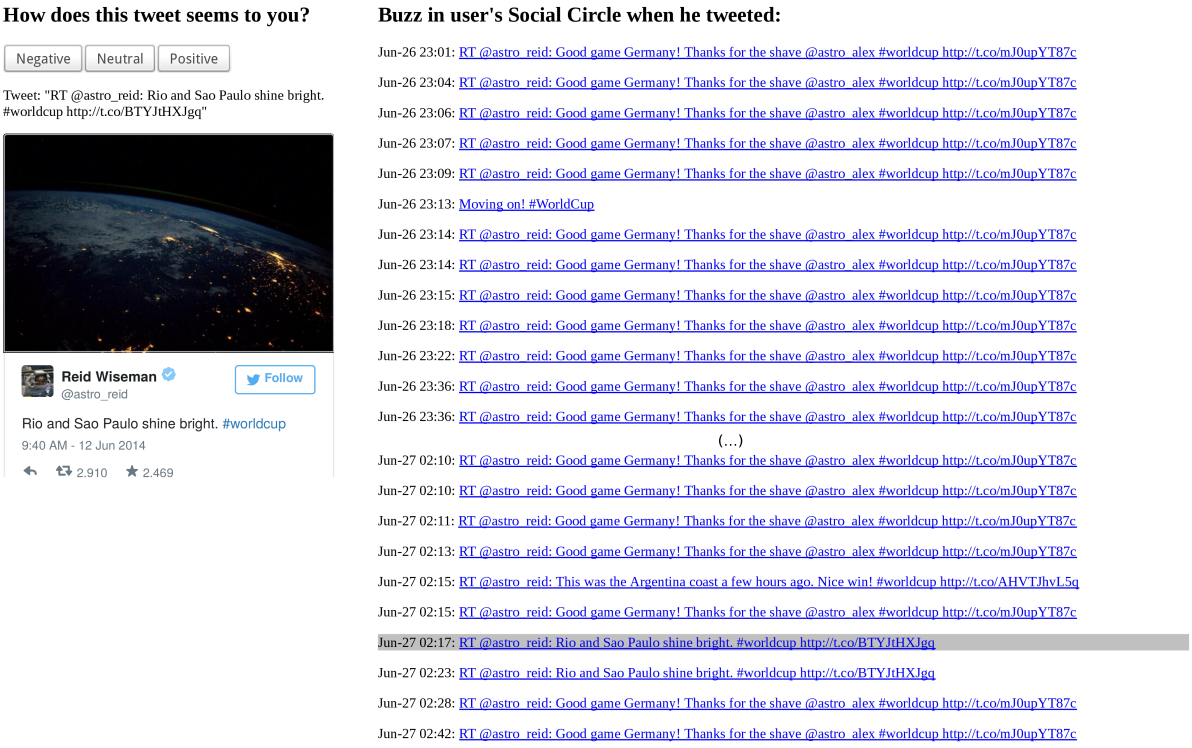


Figure C.2: Classification environment for human-coders evaluate the sentiment polarity of the displayed tweet as positive, neutral, or negative.