SENTIMENT ANALYSIS ON MULTIPOLARIZED SOCIAL NETWORKS

PEDRO HENRIQUE CALAIS GUERRA

SENTIMENT ANALYSIS ON MULTIPOLARIZED SOCIAL NETWORKS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais - Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: WAGNER MEIRA JÚNIOR

Belo Horizonte

Agosto de 2015

Guerra, Pedro Henrique Calais

G934s Sentiment Analysis on Multipolarized Social Networks / Pedro Henrique Calais Guerra. — Belo Horizonte, 2015
xvi, 107 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais - Departamento de Ciência da Computação
Orientador: Wagner Meira Júnior

1. Computação - Teses. 2. Mineração de Dados - Teses. 3. Redes de relações sociais - Teses. 4. Polaridade (Psicologia). I. Orientador. II. Título.

CDU 519.6*73(043

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Sentiment analysis on multipolarized social networks

PEDRO HENRIQUE CALAIS GUERRA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. DANIEL RATTON FIGUEIREDO
COPPE - UFRJ

DR. DANIELE QUERCIA
University of Cambridge

PROF. RENATO MARTINS ASSUNÇÃO
Departamento de Ciência da Computação - UFMG

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 24 de agosto de 2015.

Agradecimentos

This work would not be possible without the help of many people.

I especially thank prof. Wagner Meira Jr. for being my advisor not only during Ph.D, but since when I joined his research lab, in 2003, while still an undergraduate student. Wagner's support during more than 10 years was of paramount importance to my academic life.

I also thank prof. Claire Cardie and prof. Robert Kleinberg for hosting me during my visiting research period in Cornell University during October 2012 – July 2013. My time at Cornell was crucial to improve the work!

I cannot forget UOL, Google and CNPq for providing financial support during the development of my dissertation.

I also thank all colleagues from Speed lab for all the support and funny moments.

“You can have your own opinions, but you can’t have your own facts. Truth is not a democracy.”

(Ricky Gervais)

Resumo

Uma significativa fração de debates em mídias sociais está concentrado em temas que induzem polarização na sociedade – que é o processo pelo qual um grupo social se divide em dos sub-grupos com visões opostas sobre um tema. Observamos debates polarizados em uma grande gama de temas amplos e relevantes para a sociedade, como Política, esportes e políticas públicas. Nesta tese de doutorado, desenvolvemos contribuições em três direções:

1. Primeiramente, medimos a intensidade da polarização em redes sociais online que debatem um tema específico. Em particular demonstramos que a métrica de ciência de rede mais coumumente empregada para este tipo de análise (modularidade) não é adequada para discriminar polarização da ausência de polarização; ebtão propomos e avaliamos duas métricas adicionais baseadas na estrutura da rede que, como demonstraremos, capturam mais precisamente o fenômenos social de polarização.

2. Oferecemos novos métodos para processar e interpretar opiniões expressadas em discussões polarizadas online a partir do emprego de teorias bem-estabelecidas na literatura de psicologia social que descrevem como as pessoas formam as suas opiniões. Usamos estas teorias como fundações para novas sinais que habilitam métodos de análise de sentimento em cenÃ¡rios em que as opiniões chegam na forma de fluxos sociais – um fluxo de opiniões evolutivo e dinâmico.

3. A terceira contribuição está relacionada ao fato de que, em muitos domínios, mais do que dois lados estão em conflito em relação a um tópico, como em sistemas políticos multipartidários. Diferentemente do caso clássico de bipolarização, em redes sociais multipolarizadas observamos relações mais complexas, além da dualidade concordância e antagonismo. Além de demonstrar as inconsistências que não são percebidas em análises de redes sociais bipolarizadas, propomos um algoritmo que infere relações de antagonismo entre comunidades neste cenário.

Palavras-chave: Polarização, Análise de Sentimento, Mineração de Grafos, Mí dias Sociais.

Abstract

A significant fraction of debate in social media is concentrated on issues that induce *polarization* in the society – the process whereby a social group is divided into two opposing sub-groups having conflicting viewpoints. We witness polarized debate in a wide range of broad and relevant topics for society, such as Politics, Sports and public policies. In this P.hD. dissertation, we develop contributions on three main directions that analyze and make sense of "online battles" fought on social media networks over polarizing topics:

1. First, we **measure the strength of polarization** on (online) social networks with respect to a given topic discussion. We demonstrate that the current network science metric widely used to measure polarization (the well-known *modularity*) is not well suited to discriminate between polarization and absence of polarization; we then propose and evaluate two additional metrics based on the social network structure that, as we will demonstrate, capture more accurately the social phenomena of polarization.

2. Second, we **offer new methods of processing and interpreting opinions** expressed on online polarized discussions by uncovering from the social psychology literature a collection of well-established social theories that describe how people form their opinions on polarized discussions. We use these theories as foundations for new signals that enable sentiment analysis methods to operate on the classification of sentiment on opinions that arrive in the form of *social streams* – an **evolving**, **bursty** and **time-changing** flow of opinions.

3. Our third contribution is related to the observation that, on many domains, **we have more than two viewpoints in conflict with respect to a topic**, as on multipartisan political systems. Differently from the classical case of bipolarization of opinions, in multipolarized social networks we observe more complex relationships among sides, rather than the duality support/antagonism. In addition to highlighting inconsistencies that are hidden on bipolarized network analyses, we propose an algorithm that infers antagonism relationships among social communities in such a setting.

Palavras-chave: Polarization, Sentiment Analysis, Graph Mining, Social Media.

Sumário

Capítulo 1

Introduction

In Social Sciences, polarization is the social process whereby a social or political group is divided into two opposing sub-groups having conflicting and contrasting positions, goals and viewpoints, with few individuals remaining neutral or holding an intermediate position [149; 74]. A typical domain where polarization is witnessed is Politics [163; 43], although a range of other issues are known to induce in the society a divisive debate that often makes a fraction of people to have very extreme opinions, such as global warming [112], gun control, same-sex marriage, abortion [121; 73] and even religion [31; 164].

According to this sociological perspective, polarization can be formally understood as a state that "refers to the extent to which opinions on an issue are opposed in relation to some theoretical maximum", and, as a process, it is the increase in such opposition over time [121; 132]. A typical sign that polarization is playing a role in a society with regard to an issue is when opinions become more extreme over time even after opposing sides examine the *same* evidence [175], as demonstrated on a classical experiment in the 70's, where people against and in favor of death penalty have become *more* convinced of their conflicting positions after reading the *same* essay on the topic [109]. In that direction, a common approach to reason about such setting is to model the polarization phenomenon with Bayesian probabilities [43; 17]: the previous belief each group or individual has on a topic is the *prior*, and, if the updated beliefs of the opposing groups become more divergent after both examine the same evidence, then it is likely that polarization is happening.

Blogs, microblogs and online social networks are now the primary medium used by people to express their opinions about all the sort of "buzz" topics that pop up daily on news media [96; 27; 76; 20], especially those that polarize the society. As an example, see in Figure 1.1 how the U.S. political blogosphere divides itself into two polarizing, antagonistic groups (blue group = Democrats, red group = Republicans) that form around the online discussion of political topics.

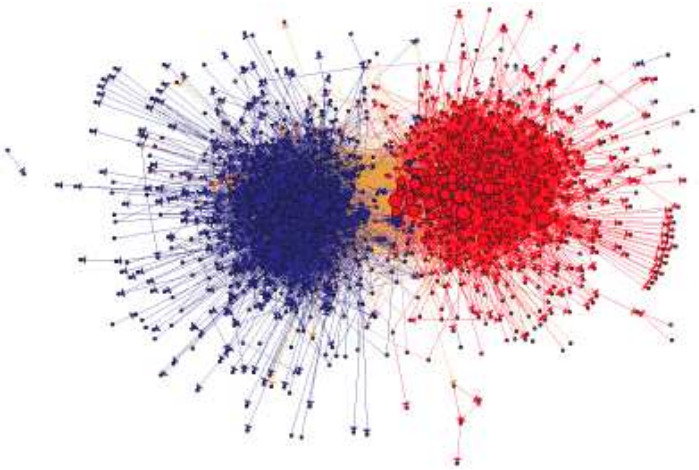


Figura 1.1. U.S. Political Blogosphere in 2004, showing the division of the network into Democrats and Republicans. Nodes are blogs and edges represent citations between two blogs (figure extracted from [3]).

Social and computer scientists are paying increasing attention to such polarizing online discussions, seeking for patterns that unveil the dynamics of online debate and the bursts of opinionated content generated in reaction to real-life events. Studying polarization as a social phenomena that takes place on social media platforms is useful for at least three reasons:

1. It is a relevant issue from the sociological point of view, since polarization causes segregation and political conflict in the society, as a consequence of the increase of extreme opinions over time and the high degree of bias of polarized opinions [132; 121]. In a recent study by Pew Research Center, polarized discussions have been identified as one of the top 6 most common conversational structures in Twitter [144].
2. Polarization may be a key information for tasks such as opinion/sentiment analysis. A biased opinion holder is likely to keep the same, extreme opinion over time, and the knowledge of the opinion holder bias in a discussion an opinion holder is (in favor or against an issue) can help predict the polarity of his/her opinions expressed on written text [150], as we will later demonstrate in this dissertation.
3. In polarized debate, the strong bias of opinions suggests that they should not be taken into consideration without considering *who* is issuing the opinion [161]. In other words, two equivalent opinions may have different interpretations and impact if issued by people from opposite sides, and new opinion mining tasks may apply, such as monitoring how many people changed their previous viewpoint over a topic.

This dissertation focuses on the analysis of “online battles” fought on social media networks over polarizing, polemic issues. We develop contributions on three main directions:

1. First, we seek to accurately *measure* the strength of polarization on (online) social networks with respect to a given topic discussion. We demonstrate that the current network science metric widely used to measure polarization (the well-known modularity [125]) is not well suited to discriminate between polarization and absence of polarization; we then propose and evaluate two additional structural network metrics based on the social network structure that, as we will demonstrate, capture more accurately the social phenomena of polarization.

   Sociologists usually resort to polls and elections data to assess the presence of extreme opinions on the public opinion [1]. When information on the relationships among people is available, polarization is commonly accepted as an ongoing phenomenon if people can be divided into highly cohesive communities (as on Figure 1.1); each community represents a distinct position or preference: liberal versus conservative parties, pro-gun and anti-gun voices, for instance. In fact, the segregation of people into groups is a remarkable characteristic of social networks induced by polarized debate as an immediate consequence of the homophily principle, which states that people with similar beliefs and opinions tend to establish social ties with higher probability [113; 175]. As a general concept, a community is cohesive if both the internal connectivity of the group is high and also the connectivity of members of the group with members from outside the community is low. Group cohesion is usually measured through community quality metrics such as *conductance* [79] and *modularity* [125]. For instance, [181] and [163] argue that modularity may be used to study partisan polarization in U.S. Congress. On the online world, modularity has been used as evidence of segregation between political groups in a diversity of online media such as blogs [3] and Twitter [33; 107].

   However, these works analyze contexts and domains which are previously known and expected to induce polarization – in particular, Politics. As a consequence, they do not examine networks from non-polarized domains, and it remains an open question what are the necessary *and* sufficient conditions for polarization between groups of individuals, in terms of the structure of the induced social network. In order to precisely understand how polarization affects the social network structure, we need to inspect *both* polarized and non-polarized domains, thus avoiding the "sampling bias" of examining only highly-polarized networks. Previous research provides strong evidence that the existence of highly modular groups is a *necessary* condition in order to observe polarization; our work contributes to the better understanding of how polarization affects the structure of social networks by performing a systematic comparison between both polarized and non-polarized networks. Instead of using modularity, we propose two new measures of polarization based on the analysis of the structure of the boundary

of two potentially polarizing groups, which we will detail and evaluate in Chapter 2.

2. Our second major contribution in this dissertation is on the field of *sentiment analysis* of polarized online discussions. In Computer Science, sentiment analysis (or opinion mining) is the set of techniques, algorithms and models that combines Data Mining, Machine Learning, Linguistics, Natural Language Processing (NLP) and Text Analytics whose goal is to analyze text fragments and determine the attitude, belief, emotion, opinion, evaluation or sentiment of a speaker or a writer with respect to some topic or entity [155; 131; 76; 36; 148]. In its simpler form, sentiment analysis is a text classification task where the goal is to classify textual content into classes $\{positive, negative\}$ regarding an entity of interest. Although still young and perhaps still lacking reliability [105], research on sentiment analysis has enjoyed relative success when applied to static and well-controlled scenarios, specially analysis of reviews of products and services on e-commerce sites. Most works report accuracy rates as high as 80% [131; 155], which is acceptable given the challenges in making machines understand free text.

   This work offers new methods of processing and interpreting opinions expressed on online polarized debate by uncovering from the social psychology literature well-established theories that describe how people form and express their opinions on polarized discussions. We embed such theories on sentiment analysis models, and our ultimate goal is to perform sentiment analysis on polarized discussions that arrive in the form of *social streams* – an **evolving**, **bursty** and **time-changing** flow of opinions, that poses challenges for data mining and machine learning research [180]. In terms of dynamicity and evolving nature, discussion on general/broad topics has a very different characteristic when compared to the classical product-review scenario, a "static" domain where we do not expect the vocabulary to be dynamic; customers are expected to comment on the features of cameras, or on the quality of the soundtrack and actors of a movie. From the machine learning standpoint, sentiment analysis (and, more generally, classification and prediction tasks associated with opinions) over social stream content is particularly challenging for three main reasons:

   a) **Challenge 1 – Lack of Labeled Data:** Traditional sentiment/opinion analysis algorithms have been designed having in mind static and well-controlled scenarios that target analysis of reviews of products and services [131; 155]. In those scenarios, even pre-defined lists of positive and negative words (i.e., lexicons) have been quite sucessful, but domain-independent lexicons tend to have low coverage when applied to specialized domains [157] who have particular idiosyncrasies of dialect and subtleties in expressing opinions on them also limit this approach [12; 10]. In case the vocabulary is not known *a priori*, supervised learning algorithms which

learn from labeled examples can also be applicable [155]. However, even supervised learning approaches are hardly applicable on social streams, due to the cost of acquiring labels. This is a consequence of the high volume and sparsity of the social stream data flow, making the acquisition of vast amounts of labeled content unfeasible, compromising the potential of typical supervised learning strategies, justifying the development of new approaches.

b) **Challenge 2 – Nonstationary Data:** Since social streams reflect the buzz of the real world, they exhibit an inherent *dynamic* nature [157]. Real-time sentiment analysis needs to deal with textual data that exhibits significant *concept drift* and non-stationary distribution, which degrades sentiment classification quality over time. For instance, at least 50 different high-volume discussion threads arose in social media systems during the U.S. 2008 Elections, as shown in Figure 1.2 provided by [96] (such as the famous "lipstick on a pig" discussion); during the 2010 Brazilian Presidential Elections a scandal involving a close assistant of one of the candidates was unveiled in the middle of the electoral process, unleashing a large volume of unpredicted negative comments. In Sports scenario, changes in sentiment also occur in sudden and unexpected bursts; in a few minutes, a positive sentiment of the fans can be destroyed by a sequence of good actions from the adversary team.

c) **Challenge 3 – Lack of reliability:** although a promising and advancing area of study, sentiment analysis is still far from being a task which is considered reliable [Metaxas et al.], specially due to the intricacies of human language, what discourages companies to fully adopt opinion analysis in their web systems. The informal and diverse language of social media, in constrast to the well-formed data obtained from news articles, makes the applicability of traditional sentiment analysis techniques on social media data less attractive [157].

Note that the text-based sentiment classification task we are interested simultaneously i) lacks the support of labeled examples and ii) needs to deal with nonstationary data, what turns the content classification problem even harder since some state-of-the-art solutions from one challenge involves assumptions that are not true due to the other challenge. The state-of-the art solution for **Challenge 1** is *semi-supervised learning*, a field of machine learning that makes use of both labeled and unlabeled data for model generation [183]. However, due to **Challenge 2**, the usefulness of the few available labeled examples can be very limited since they can become quickly outdated. On the other hand, the traditional solution for **Challenge 2** is to incrementally update the model through fresh, recently-acquired labels that are provided by the stream [58; 167]

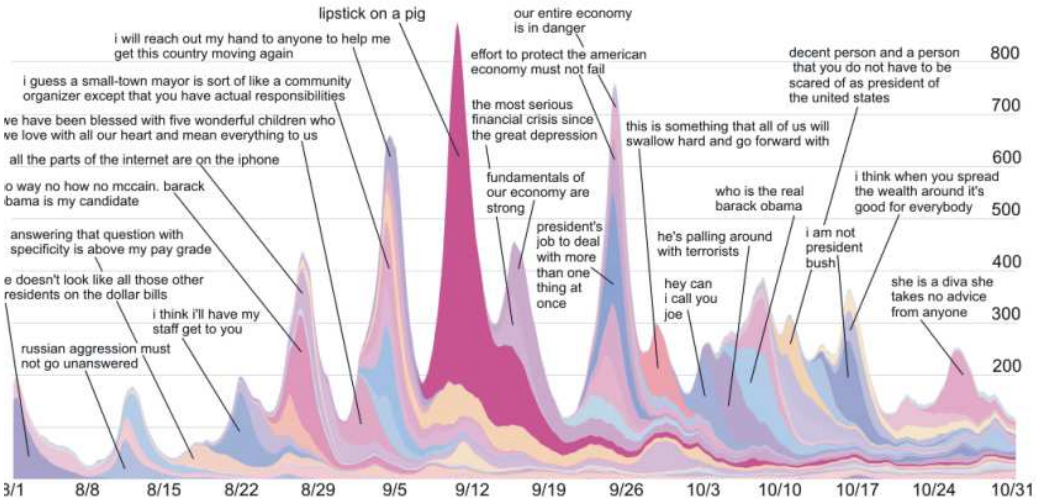


Figura 1.2. Debate on news media and social media tends to cycle through new, emerging issues that arise due new events that occur, as during 2008 U.S. Elections [96]. Processing and interpreting opinions on this dynamic scenario is a challenge.

but such solution may not be feasible due to **Challenge 1**.

The dominant approach for sentiment classification (a.k.a. sentiment analysis) is to treat it as a pure text classification problem. Different text classification algorithms have been applied to learn from word co-occurences and linguistic features to determine the sentiment contained in documents [105]. Moreover, these algorithms have been used in conjunction with pre-defined lexicons to assess sentiment in political and movie review blogs [155]. These approaches do not solve properly the 3 challenges we discussed in the previous section, since terms and expressions associated with debates on polarizing topics change according to real world events.

Polarization and Biased Opinions. The key point that we aim to explore in this dissertation is that polarization is related to strong bias on opinion holders' opinions. While textual content may be influenced by external factors, such as new terms that enter a topic discussion, user biases are less prone to external perturbations, unless users actually change their opinion, which is usually a relatively longer process. Thus, opinion holder bias patterns can potentially give us the capability to deal with the challenges of lack of labeled textual data to support learning algorithms and the unpredictable directions discussions can take, because, as we will show, biased users act as a reliable source of labels that indicate if a content carries a positive or negative sentiment and, since biased opinion holders seldom change their global viewpoint regarding a topic, their activity allows us to deal with the unpredictability of new content that arise. Such capability comes from the **social** and **temporal** contexts that polarized debate induce on opinion

holders.

As we will detail in Chapter 3, a range of cognitive biases studied by social psychologists (e.g., *confirmation bias*, *hindsight bias* and *self-reporting bias*) explain how human reasoning works and how humans tend to favor opinions they *already* have [134], supporting our claim that correlation between opinion holders and opinion should be explored in the context of polarized debates. Furthermore, the psycho-social background we exploit here supports our search for *predictable* behaviors, which is key to address Challenge 3. It is also worth noting that the ambiguity and lack of context inherent to social media messages make hard to textual-based analysis techniques to perform good predictions.

The social psychology background is the first step of our opinion analysis framework shown in Figure 1.3. In the second phase of our framework, we look forward detecting and inferring such cognitive biases from social interactions available from social media data (e.g., Facebook, Twitter, among others). We then extract network structures and regularities that are meaningful for the sake of finding the individual opinions in the network regarding the considered topic.

Finally, we intend to combine the learned patterns from polarized networks to measure and track opinions in the context of interest. We combine the learned patterns from the polarized networks structure to measure and track opinions on the polarized context of interest. We do that by providing a transfer learning framework [130] which converts opinion holder patterns to content polarity predictions by exploring the correlation between opinion holders and opinions on polarized discussions. Transfer learning is a field of machine learning which focuses on using knowledge acquiring from one to task to solve a different, but related task [130]. Such approach is specially useful when a target task τ_t is hard to solve, but it may benefit from knowledge obtained from a similar source task τ_s.

3. Our third contribution is the observation that, on many domains, we have **more than two viewpoints** in conflict with respect to a topic: think, for example, on multipartisan political systems (as in Brazil), sports competitions (32 countries take part on the Soccer World Cup; almost 200 play on on the Olympics) and TV Reality Shows (on *Big Brother*, groups of supporters are formed around each of the 10 participants). In such scenarios, we observe more complex relationships among sides (such as indifference), rather than the duality support/antagonism. For instance, we have found so far that fans of soccer clubs in Brazil which share a strong are *closer* to each other in an endorsement network (for example, a network connecting users by their *retweets*) than to fans of other clubs. This apparent paradox occurs because other effects such as the impor-

tance that one side plays to the other overlaps with other motivations for establishing (or not) social ties and interactions, and one can broadcast a message they disagree with for a variety of reasons: sarcasm, irony, to show their audience how wrong a contrary opinion is etc.

Our innovation in this step includes avoiding the ambiguity in the signal provided solely by retweets and replies by exploiting a more discriminative signal to detect antagonism: the *lack* of interaction between some specific sets of users and messages, in particular, we exploit the *negative implicit feedback* carried on absence of interaction between sets of users and message where interaction were expected if users viewed such messages as positive with respect to theit viewpoint.

This dissertation is organized as follows:

1. In Chapter 2, we present our novel metrics designed to quantify evidence of polarization based on the structural characteristics of a social network built around a topic discussion;
2. In Chapter 3, we present our method that infer the bias of opinions holders regarding a topic based on their social interactions;
3. In Chapter 4, we demonstrate how social theories can be embedded on sentiment analysis algorithms in order to perform sentiment analysis on social streams of polarized discussions.
4. Finally, in Chapter 5 we summarize our contributions and discuss future work.

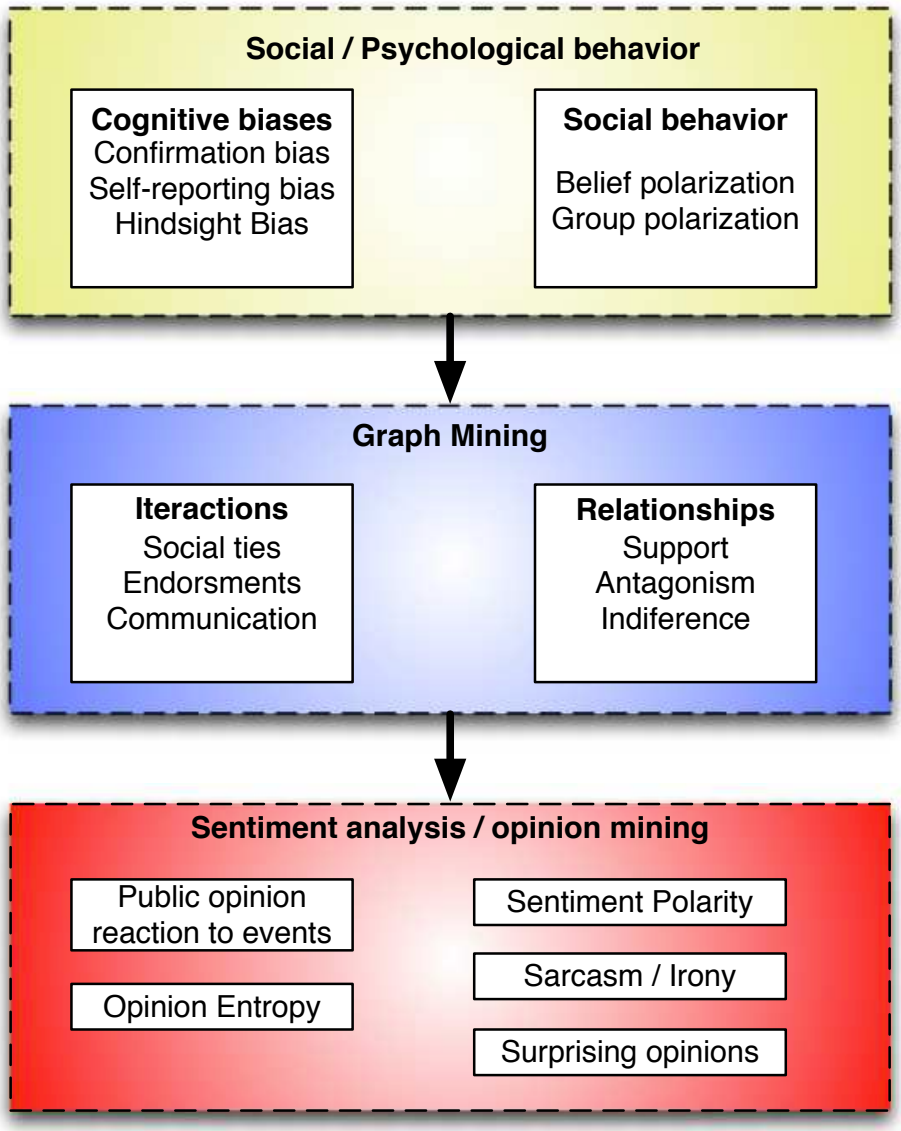


Figura 1.3. Three-layer framework comprising 1) opinion holders' behavior, 2) network structure and dynamics, and 3) inferring of opinions and opinion dynamics.

Capítulo 2

Measuring Polarization on Social Networks

Given the relevance of the contexts in which polarization is witnessed, many works from Computer Science (and, more specifically, Social Network Analysis) have investigated online social networks induced by polarized debate, specially in the political domain [3; 107]. In general, the Computer Science literature assumes (either implicitily or explicitly) that a social network is *polarized* if nodes can be partitioned into two highly cohesive subgroups, reflecting, possibly, two contrasting viewpoints. In particular, the well-known community quality metric known as *modularity* [125] is commonly used to measure the level of segregation of two particular groups; such as democrats vs republicans, people in favor or against abortion etc. A network segmentation with high modularity indicates that the social graph may be divided into clusters having many internal connections among nodes and few connections to the other group, what is widely accepted as an indication of polarization [33].

Our main claim in this Chapter is that, although we acknowledge that modularity is correlated to the social phenomena of polarization, and highly modular networks are certainly linked with an increased likelihood of polarization of positions expressed by users who are part of the network, modularity is not accurate in distinguishing presence from absence of polarization. We draw this observation from the fact that it is not clear how much modularity is "enough" to state that a social network is polarized. For instance, people may be divided into those that like basketball and those that like football, even though there is no notion of opposition among the two groups – they are just two different preferences which are not mutually exclusive, since some individuals can be practitioners of both sports and thus belong to both communities. Although the existence of two segregated social groups is certainly a *necessary* condition for polarization, the modularity measured for any network divided into two cohesive communities will be a value different from zero, even if no polarization *at all*

is present among nodes.

Our goals in this Chapter are twofold. First, we perform a systematic comparison of social media networks emerging from both polarized and non-polarized contexts, by collecting a diversity of social networks from social media systems such as Twitter, Facebook and blogs. Our goal is to avoid the bias of current works, which focus on networks from domains that are previously known to be polarized, specially Politics. We then identify communities in these networks and verify that their modularity measure is not sufficiently clear to state that polarization is an ongoing phenomenon or not; although polarized social networks tend to be more modular than non-polarized networks, the determination of a threshold of polarization is a challenging task that depends on factors such as social media platform and nature of interactions.

Motivated by this observation, we focus on the following question: given that polarization is recognized as a strong, remarkable sociological phenomena, are there structural patterns which better capture the differences between polarized and non-polarized networks, rather than the level of modularity between communities? We propose an analysis of the boundary between the two potentially polarizing communities – the portion of the social graph comprising nodes from one community which link to one or more nodes of the other community. Our hypothesis is that, in such community boundaries, one group unveils what they "think" about the other group, and thus it is the place where we should seek for evidences of antagonism. Our metric considers a null model of polarization that assumes that, on a non-polarized network, cross-group interactions established by member of a community boundary should be at least as frequent as interactions with internal nodes on the community. The model considers nodes' likelihood into connecting to users which belong to the other (potentially opposing) group, in comparison to the likelihood of connecting to members from its own group.

We also empirically demonstrate that polarized and non-polarized social networks tend to differ according to another structural property: the concentration of popular (high-degree) nodes not belonging to community boundaries. On non-polarized contexts, we observed a concentration of popular nodes along the boundary, since the sharing of similarities between members of the boundary increase the popularity of such nodes (e.g., users that like both football *and* basketball). On the other hand, we found that polarized networks tend to have a lower concentration of popular nodes in the boundary, since the antagonism between both sides decrease the likelihood of existence of nodes that are popular in both groups.

To show the applicability of our findings on the interpretation of opinions expressed on social media, we employ our metrics to perform an analysis of opinions expressed on Twitter on the gun control issue in the United States. We demonstrate that our metrics based on community boundaries are a useful complement to the traditional modularity measure in

helping to understand how the structure of a social network links with the viewpoints and opinions expressed in online social enviroments.

This Chapter is organized as follows. In Section 2.1 we evaluate the modularity of a range of polarized and non-polarized networks. We then propose a new metric to measure polarization based on community boundaries, in Section 2.2. In Section 2.3, we employ our metric to understand opinions expressed on Twitter on the gun control issue on the United States. Next, we compare polarized and non-polarized networks in terms of another structural property – concentration of popular nodes in the boundary – in Section 2.4.

2.1 Modularity as a Measure of Polarization

Modularity is widely used as a measure of polarization of a social network: for instance, [181] and [163] argue that modularity may be used to study partisan polarization in U.S. Congress. On the online world, modularity has been used as evidence of segregation between political groups in a diversity of online media such as blogs [3] and Twitter [33; 107]. Here, we consider existing and publicly available social networks and additional networks we collected from Facebook and Twitter in order to compare the structural characteristic of social networks with varying degrees of polarization, including total absence of polarization. We use the following social networks:

1. **University Friendships' Network:** This social network comprises the social relationships established on Facebook by professors, undergraduate and graduate students of a large department at a Brazilian University.

2. **Brazilian Soccer Supporters:** We collected mentions, on Twitter, to two of the most popular soccer teams in Brazil – Cruzeiro and Atletico Mineiro, known by being the fiercest rivals in the country. Nodes are Twitter users, and a direct edge connects users involved in any *retweet*. A *retweet* usually means an endorsement [29], and thus it is a good evidence of sharing of similar viewpoints between two individuals (we will relax this definition on Chapter 3).

3. **New York City Sports Teams:** We collected mentions, on Twitter, to two sports teams hosted in New York City: New York Giants (football) and New York Knicks (basketball). The network is induced by *retweets*; we restrict the network to nodes that mentioned both teams at least once, to guarantee that we are taking into account only users who are interested in both teams. Note that, differently from the previous network, we do not expect polarization here, since the two potential communities represent supporters of teams from different sports.

4. **Karate's Club:** This is a social network of friendship ties established between 34 members of a karate club at a U.S. university in the 1970s, and the emergence of two communities was a result of a disagreement developed between the administrator of the club and the club's instructor, which ultimately resulted in the instructor's leaving and starting a new club, taking about half of the original club's members with him [179; 46].

5. **2004 U.S. Political Blogosphere:** This dataset was among the first that showed that political blogs on the U.S. are divided into two dense communities – representing liberals and conservatives [3]. Directed edges are links between two blogs.

6. **Gun Control:** We collected tweets mentioning gun control issues since the shootings on Sandy Hook Elementary School in Newtown, Connecticut, on December 14, 2012. We considered the following keywords to collect data: `gun control`, `guns`, `mass shootings` and `NRA`[1]. As in other networks obtained from Twitter, users are linked through *retweets*.

Note that all the aforementioned networks have a semantic unicity, in the sense that users interacting and expressing opinions are restricted to a single domain or topic. In Table 2.1 we provide a summary of the main characteristics of these networks, including number of nodes and edges. For each network, we split nodes into communities, in order to assess the structural patterns that arise from the segmentation of the graph into groups. In the case of the networks `University`, `Brazilian-Soccer`, `Political-Blogs` and `Gun-Control`, we have run the community detection algorithm from [24], a simple modularity maximization approach provided by the *Gephi*[2] software package. In the case of the network `NYC-Teams`, we separated users into the community of NY Giants or NY Knicks according to the number of hashtags each user posted referring to each team. For network `Karate-Club`, we employed the ground-truth separation provided by [46]. The gun control debate graph was divided into three large communities, and we leave its analysis to Section 2.3, after we introduce our novel polarization metric.

Tabela 2.1. General Description of Social Networks and derived communities.

network	media	# nodes	edge type	# edges	communities	modularity Q
1 - NYC Teams	Twitter	19,585	directed	201,691	NY Giants fans and NY Knicks fans	0.15
2 - University	Facebook	133	undirected	2,241	graduate and undergraduate students	0.24
3 - Karate's Club	friendships	34	undirected	78	followers of instructors 1 and 2	0.35
4 - Brazilian Soccer Teams	Twitter	27,415	directed	156,489	Cruzeiro and Atletico fans	0.39
5 - US Political Blogs	blogs	1,224	directed	16,715	liberals and conservatives	0.42
6 - Gun Control	Twitter	61,740	directed	342,449	analyzed in Section 2.3	–

[1] NRA is the National Rifle Association.
[2] *Gephi* is available at `http://www.gephi.org`.

For each pair of communities, we calculate modularity Q. The modularity of a network quantifies the extent, relative to a random network, to which vertices cluster into community groups, and the higher its value, more modular the network is [125]. Modularity is traditionally formulated as Equation 1; m is the number of edges, A is the adjacency matrix, k_i and k_j are node degrees and $s_i s_j = 1$ if nodes i and j belong to the same community and -1 otherwise. Values of Q obtained for the datasets we consider in this analysis are shown in Table 2.1.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \frac{s_i s_j + 1}{2} \tag{2.1}$$

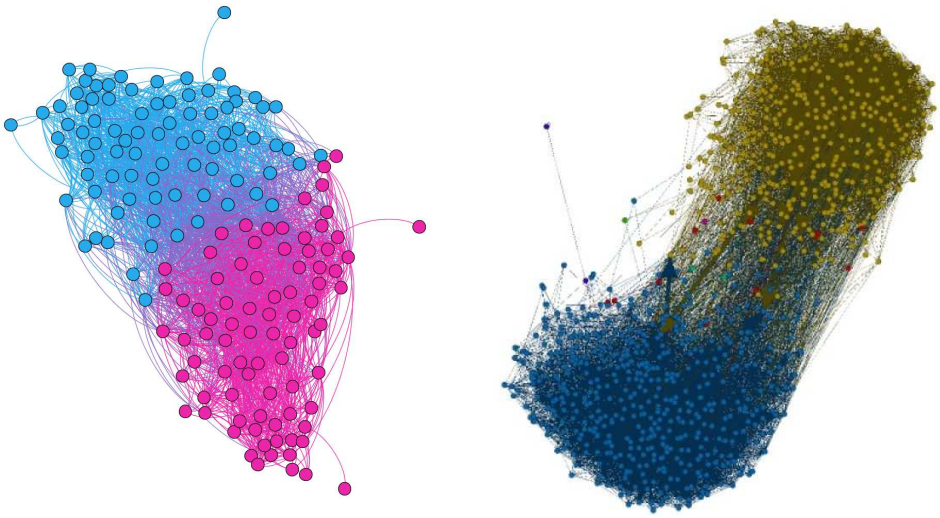
We first observe that networks induced by domains for which we expect polarization (networks 3, 4 and 5) exhibit a high measure of modularity when compared to networks 1 and 2. This observation is in accordance with previous works that associate high modularity to polarization [33]. However, we point out three drawbacks on mapping modularity to the sociological behavior of polarization:

1. On communities that arise from contexts where we do not expect polarization, the modularity value is still a positive, moderate value, as in the case of `University` and `NYC-Teams` networks. Modularity for the `University` network is 0.24 (shown in Figure 2.1(a)), what suggests a network less polarized than `Political-Blogs`, which exhibits a modularity value of 0.42 (Figure 3.1(a)). However, from the sociological standpoint, we expect to observe *any* (or little) antagonism at all between undergraduate and graduate students.

2. The direct mapping of modularity values into degrees of polarization shows some inconsistencies when we compare modularity measures obtained by independent researches working with different data. [181], for instance, have found modularity values not higher than 0.18 from the examination of networks induced by voting agreement on the U.S. Congress. Although the authors' goal is to evaluate the increase of modularity over time to conclude that polarization was rising among politicians over the decades, the maximum modularity measure they found is just 0.01 higher than the value that [33] found to conclude that $Q = 0.17$ is not associated with an evident community structure on a communication network in Twitter. In previous researches, modularity is used more to *confirm* an early suspection of polarization, rather than find whether polarization exists or not in an unknown domain.

3. Modularity has a known resolution limit problem caused by the fact that its null model assumes that each node may connect to any other node of the network, what is not realistic for large graphs [62]. Therefore, comparing the modularity value across different

networks is not a good practice if the graphs' size are very different [48], which is the case of the graphs compared in Table 2.1.



(a) Network of Facebook Friends from an University Department

(b) 2004 U.S. Political Blogosphere

Figura 2.1. Two social graphs showing a non-polarized network (Facebook Friends) and a polarized network (Politics). Although the political network is more modular, it is not clear what is the minimum level of modularity associated with the sociological phenomenon of polarization.

The conceptual gap between the modularity measure and the sociological behavior of polarization, evidenced on these extreme cases, limits the understanding of networks and contexts where it is less clear whether polarization is taking place. In the next section, we will provide details about a novel structural pattern we propose, in order to better capture the presence and absence of polarization in communities formed around a given domain or topic of discussion.

2.2 Measuring Polarization on Community Boundaries

It is known that a significant portion of the structure of a social network is affected by the context and the behavior of the nodes [46]. Behavioral patterns such as homophily [113], social influence [50] and social balance [69] directly affect the likelihood that specific pairs of users will establish a tie in a social environment. Since polarization is a strong, remarkable sociological phenomenon, we expect that a social network embedded in such a context of

opposing and conflicting relationships will induce structural patterns which are not observed on general, non-polarized networks.

The link between high modularity and polarization carries the implicit assumption that the absence of positive interactions between nodes (e.g., message sharing, *retweets* and friendship ties) is a sign of *antagonism*, i.e., a segmentation of social groups due to opposition and clash of viewpoints. Modularity compares the internal and external connectivity of two groups G_i and G_j; it quantifies both homophily (nodes from a community establishing ties due similarity) and antagonism (nodes avoiding establishing ties with the alternate community) through the same equation, and limits the understanding of antagonism in isolation. To better understand polarization, we propose to seek for social structures that highlight the presence (or absence) of antagonism, since homophily is a pattern present both on non-polarized and polarized networks, but antagonism is expected only on the latter. Recent work show that homophily by itself do not explain polarization and additional social phenomena should be taken into account [35], we thus expect the network structure of polarized social networks to reflect such differences.

With this idea in mind, we focus our analysis on nodes that effectively interact with the (potentially) opposing group. These nodes are part of a *community boundary*, which we define, for a group/community G_i, as the subset of nodes $B_{i,j}$ that satisfies two conditions:

1. A node $v \in G_i$ has at least one edge connecting to community G_j;
2. A node $v \in G_i$ has at least one edge connecting to a member of G_i which is not connected to G_j. This is to guarantee that v is connected to nodes which do not belong to the boundary.

Equation 2.2 formally defines boundary $B_{i,j}$. In Figure 2.2 we show a toy example of a network divided into communities G_1 (dark) and G_2 (white). According to our definition, $B_{1,2} = \{b, d\}$ and $B_{2,1} = \{1, 2\}$. Note that node e does not belong to $B_{1,2}$ because it does not meet condition 2; we exclude it from the boundary because we cannot guarantee that e knows both set of nodes (internal and boundary nodes).

$$\begin{aligned} B_{i,j} = \{v_i : v_i \in G_i, \exists e_{ik} | v_k \in G_j, \\ \exists e_{ik} | (v_k \in G_i, e_{kl} | v_l \notin G_j), i \neq j\} \end{aligned} \tag{2.2}$$

Nodes from G_i which do not belong to $B_{i,j}$ are named *internal nodes* and are grouped in the set I_i, defined by Equation 2.3. In Figure 2.2, $I_1 = \{a, c\}$ and $I_2 = \{3, 4\}$.

$$I_i = G_i - B_{i,j} \tag{2.3}$$

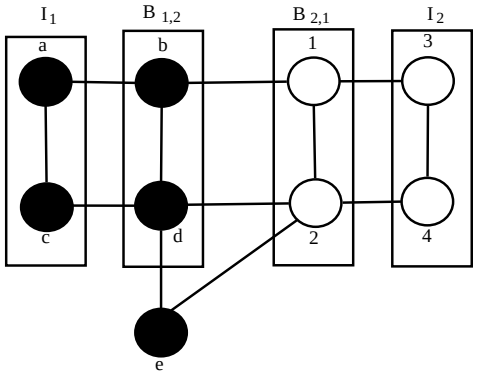


Figura 2.2. Toy example of a graph divided into two communities G_1 and G_2. Sets I_1, $B_{1,2}$, $B_{2,1}$ and I_2 are defined according to Equations 2 and 3.

We perform our analysis of polarization by analyzing the connectivity between I_i, $B_{j,i}$, I_j and $B_{i,j}$. These four sets allow us to compare nodes' choices in connecting to nodes from a very different nature. Due to condition 1, we can assess the connections of $B_{i,j}$ with $B_{j,i}$, i.e., with nodes that belong to a potentially opposing group. Due to condition 2, nodes from $B_{i,j}$ also establish contact with a set of nodes which do not connect to any member of the potentially opposing group. Nodes from I_i avoid any connection to the alternate group and restrict their connections to nodes from their own community, representing individuals that, theoretically, are very different from nodes from the other group, in the case of a polarizing domain.

We focus on I_i, I_j, $B_{i,j}$ and $B_{j,i}$ as groups that better represent the (potential) distinct nature of (potentially) polarized individuals, in comparison to the division between G_i and G_j that is analyzed by modularity. Our proposal is to compare the degree of preference of each node in $B_{1,2}$ to connect to members from I_1 or $B_{2,1}$, and of each node in $B_{2,1}$ to connect to members from I_2 or $B_{1,2}$. To perform such comparison, we define two sets of edges. The first set is E_B, which is the set of edges that connect members from G_i to members from G_j:

$$E_B = \{e_{mn} : v_m \in B_{i,j} \wedge v_n \in B_{j,i}\} \tag{2.4}$$

In Figure 2.2, $E_B = \{(b, 1), (d, 2)\}$. These edges are evidence of interaction between the two distinct groups. To contrast with these interactions, we also define E_{int} as the set of edges that connect boundary nodes to internal nodes:

$$E_{int} = \{e_{mn} : v_m \in (B_{1,2} \cup B_{2,1}) \wedge v_n \in (I_1 \cup I_2)\} \tag{2.5}$$

In the example, $E_{int} = \{(a, b), (c, d), (1, 3), (2, 4)\}$. The modularity for this community configuration is $Q = 0.30$, what indicates a reasonable level of segregation among the

two communities. However, let us examine the decisions taken by each node at the boundary in establishing their connections. Consider node b, which has a node degree $d(b) = 3$:

1. $(b, 1)$ is a cross-group edge and belongs to E_B;
2. (b, a) is an internal edge and belongs to E_{int};
3. (b, d) is neither an internal edge, nor a cross-group edge.

We consider that b did not exhibit any type of antagonism to members of the other group; since it established the *same* number of connections to $B_{2,1}$ and I_1. Note that the same reasoning is applicable to the remaining members of the boundary, d, 1 and 2. The network from Figure 2.2, according to our principle, does not exhibit polarization. Note that edges (b, d), (a, c), $(1, 2)$, $(3, 4)$ and $(e, 2)$ are intentionally not included in this evaluation, since they capture more homophily between nodes than antagonism between groups.

Equation 2.6 generalizes the comparison among the connectivity choices that nodes in $B_{i,j}$ make while connecting to members from I_i or $B_{j,i}$. For each node v belonging to the boundary B, we compute the ratio between the number of edges it has in E_{int} (which we call $d_i(v)$) and the total number of edges in E_{int} *and* edges in E_B (which we call $d_b(v)$). We compare such ratio with the following null hypothesis: each node spreads its edges equally between internal nodes and nodes from the other community. P lies in the range (-1/2,+1/2); a P value below 0 indicates not only lack of polarization, but also that nodes in the boundary are more likely to connect to the other side. Conversely, a P value greater than zero indicates that, on average, nodes on the boundary tend to connect to internal nodes rather than to nodes from the other group, indicating that antagonism is likely to be present. In the case of the communities shown on Figure 2.2, $P = 0$, since all boundary nodes established the same number of connections to internal nodes and to nodes from the alternate community.

$$P = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right] \tag{2.6}$$

P is somehow similar to *E-I Ratio* [85], a social network measure that compares the relative density of internal connections within a social group compared to the number of connections that it has to the external world. It was used in the context of organizational networks and was designed to demonstrate how informal networks crossed formal internal group structures; here we use the similar idea of comparing of the ratio of internal/external edges, but with the goal of measuring polarization in mind.

Absence of Boundary. While traditional community quality measures such as modularity are relatively high for a network comprised of two isolated communities, our polarization metric cannot be computed when $B = \emptyset$. While this case can be interpreted as a network of very high polarization, we consider that it is more reasonable to state that it is not possible

to assess polarization between two isolated communities, since it can be the case that each group does not know each other at all. The intuition here is that the hypothesis is not verifiable, since the groups do not have any interaction and we cannot guarantee that there is any polarization. It corresponds to asking whether there is polarization between human beings and extraterrestrials.

Tabela 2.2. Modularity Q and Polarization P for networks described in Table 2.1.

network	**media**	**modularity** Q	**polarization** P
1 - NYC Teams	Twitter	0.15	-0.002
2 - University	Facebook	0.24	-0.24
3 - Karate's Club	friendships	0.35	0.17
4 - Brazilian Soccer Teams	Twitter	0.39	0.20
5 - US Political Blogs	blogs	0.42	0.18

In Table 2.2 we compare values of modularity Q and polarization P for the set of datasets we consider in this work; networks are sorted according to their modularity values. Although higher values of P tend to correlate with higher values of Q, we can observe an important difference with respect to the sign of the measured values. For the network comprising supporters of New York City football and basketball teams (NY Giants and NY Knicks), our metric P detects absence of polarization ($P = -0.002$), suggesting that although fans are divided into two groups, they do not oppose each other. This is different from network 4, which comprises fans of two rival soccer teams from Brazil; in this case our metric indicates that there is, indeed, polarization among such fans ($P = 0.20$). The `University` network exhibits a negative value $P = -0.24$. This result is consistent with recent work that examine the overlap between communities in social networks and concluded that the overlap tend to be denser, in terms of number of edges, than the group themselves [172]. The boundary connects users that share common interests and background, such as supporting both NY Knicks and NY Giants or having attended high school and college together. In the case of polarized communities, such pluralistic homophily is not present.

In order to highlight the differences in the structure of large polarized and non-polarized online social networks, we compare in Figure 2.3 the node-specific values of $\frac{d_i(v)}{d_b(v)+d_i(v)} - 0.5$, which we call P_v, for each node v on the boundary of each network. The number of nodes with $P_v < 0.5$ is very limited on the polarized network of Brazilian soccer rivals, indicating their likelihood to connect to internal nodes rather than endorsing (retweeting) adversaries. Note, also, that the slope of the curve formed by points with $P_v < 0$ on the polarized network is more inclined, reflecting that nodes face resistance to connect to the boundary. We interpret such difference w.r.t. slope as a genuine manifestation of antagonism. In the curve of the non-polarized network, however, the slope before and after

$P = 0$ is roughly the same, indicating that nodes present the same likelihood to establish connections, what we interpret as a sign of absence of polarization.

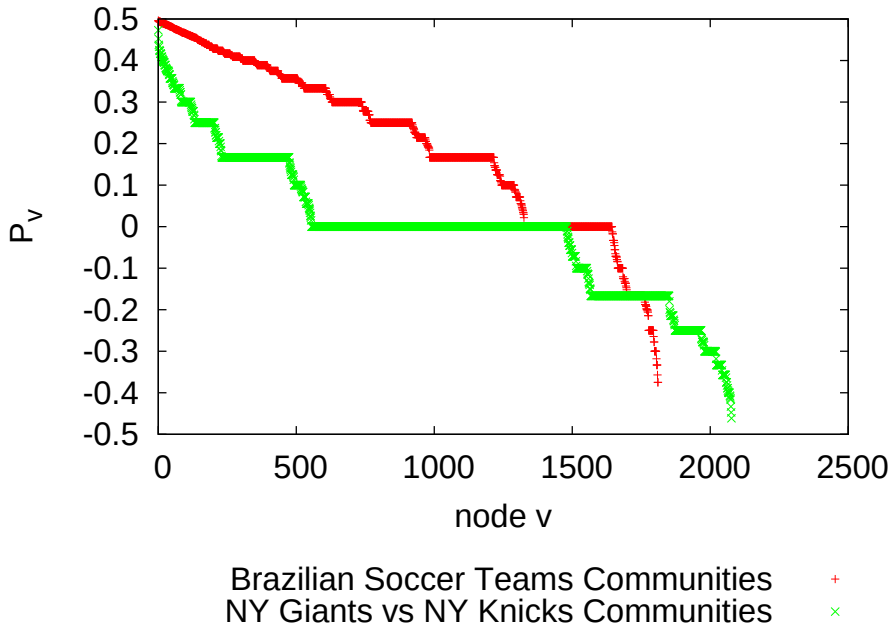


Figura 2.3. P_v for users belonging to Twitter communities debating Sports. A polarized social network is characterized by a small number of nodes that prefer cross-boundary connections ($P_v < 0$).

2.3 Opinion Analysis on the Gun Control Debate

In this section we use the polarization metric P we introduced in the last section to analyze opinions expressed on the gun control issue in Twitter. The debate around gun control laws has long history in the United States and is often present in political debates [23]. Events related to the issue, such as the shootings in the Sandy Hook Elementary School in Connecticut, on December 14, 2012, unleash bursts of strong opinions on the topic. From that date until February 10, 2013 we collected 3,816,137 tweets mentioning gun control-related keywords. Since gun control is a typically polarizing topic, we attempt to use the network structure to interpret, predict and analyze opinions expressed regarding the issue.

When plotting the social network induced by *retweets* on *Gephi* and executing the modularity maximization algorithm from [24], we got the three communities shown in Figure 2.4. We start by computing modularity Q for each of the three pairs of communities: the modularity between the leftmost group (colored in green) and the rightmost group (in yellow) is $Q = 0.47$; while modularity for communities 1 and 2 is $Q = 0.31$. Finally, the

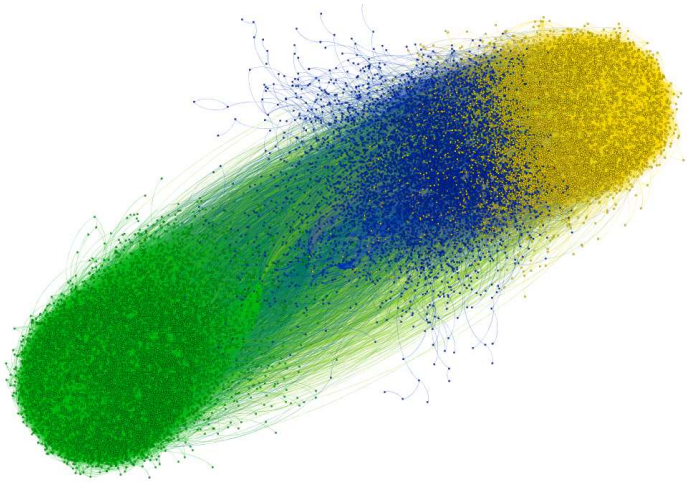


Figura 2.4. Communities obtained from gun control debate on Twitter. Nodes are users and edges represent *retweets*. From the left to the right, we refer to them as communities 1 (green), 2 (blue) and 3 (yellow).

modularity between groups 2 and 3 is $Q = 0.26$. Although we expect the most distant groups to have conflicting opinions and viewpoints, the lack of a more precise measurement of how polarization limits the understanding of the opinion sharing patterns among nodes. Does group 2 has a different, third opinion in comparison to group 3, or do they share a common viewpoint, and the division into two communities is caused by other factors? This answer is not provided by the analysis of the modularity metric by itself, because we do not know in advance whether $Q = 0.26$ is high enough to state that there is antagonism between community members, or if such threshold exists and is dependent of the social media platform or the nature of the interactions.

To gain insights on the relationships between the groups, we calculate the metric P we proposed for each pair of communities. Results are shown in Table 2.3. By analyzing the Q values, it is not immediately obvious what are the sharing and conflicting opinions between groups. However, our polarization metric P provides better clues on the opinion sharing patterns. Community 1 is predicted to be polarized with communities 2 and 3, with $P = +0.23$ and $P = +0.32$, respectively. On the other hand, our metric predicts that communities 2 and 3 have no polarization at all ($P = -0.14$). On the contrary, a P value significantly below zero means that nodes in the boundary tend to establish more cross-group connections than expected. By manual verification of a sample of the profiles of users belonging to each group, we concluded that group 1 is dominated by conservative voices, while liberals are concentrated on group 3. Group 2 is dominated by independent opinion holders.

In Figure 2.5, we plot the distribution of P_v for the boundary nodes for each pair of

Tabela 2.3. Modularity Q and Polarization P for Gun Control debate.

communities	**modularity** Q	**polarization** P
GC-1 and GC-2	0.31	+0.23
GC-1 and GC-3	0.47	+0.32
GC-2 and GC-3	0.26	-0.14

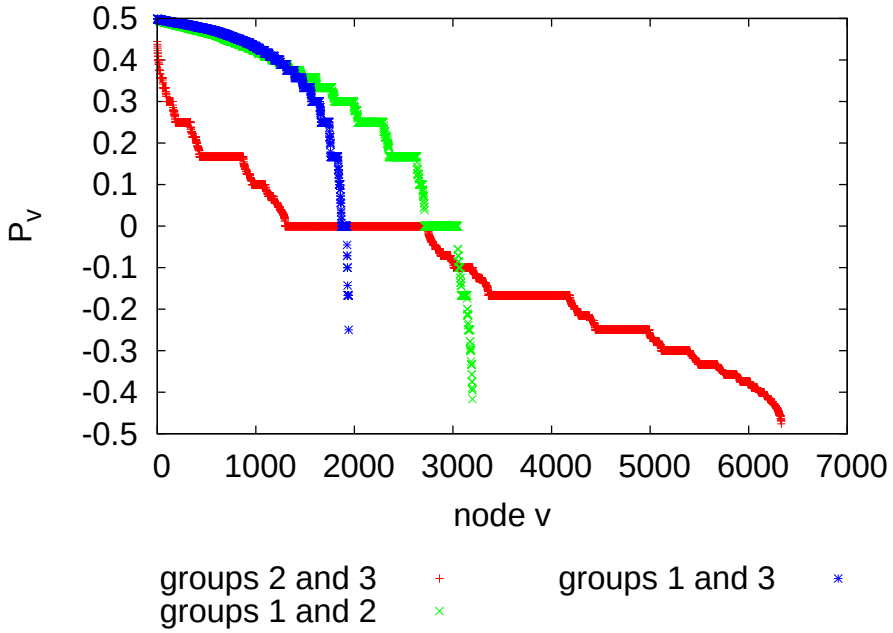


Figura 2.5. P_v for users nodes belonging to Twitter communities debating Gun Control. On the pairs of polarized communities (1–2 and 1–3), few nodes establish more connections with the alternate group than with internal nodes ($P_v < 0$).

communities. We can note a clear difference in the shape of the curve corresponding to the pair of communities 2–3 in comparison to 1–2 and 1–3: in addition to a significant number of nodes with $P_v < 0$ on the non-polarized network, the smooth transition from nodes that are more likely to connect to internal nodes from nodes that are more likely to connect to boundary nodes contrasts with the quickly decrease in the polarized curves, indicating that the boundary reduces the likelihood of connections, acting as a barrier. Moreover, the difference in modularity Q for pairs 2–3 and 1–2 is just 0.05, however their structure is fundamentally different, as Figure 2.5 shows. We believe that the different distributions we found may support the building of graph generation models that better represent polarized and non-polarized social networks.

In Table 2.4, we present some of the most popular tweets posted by members from each community comparing statistics, facts and gun regulations from three other countries – China, Australia, and Canada, in addition to the United Kingdom. Tweets from group 1 show

Tabela 2.4. Popular Tweets on Gun Control debate since Newtown Shootings, on December 14, 2012, for each community shown in Figure 2.4. Confirmation bias on polarized debate makes people focus on facts that confirm their previous opinion.

group	Twitter user	tweet	#RTs
1	@tillerylakelady	2 those of you whining about #gun control-a madman used a KNIFE to stab 20+ kids in **China** today. Its not about guns,its about mental health.	71
1	@JohnGaltTx	Since the **Australia** Gun Ban, the following increased: Armed Robberies +69%, Assaults with Guns +28%, Gun Murders +19%, Home Invasions +21%	12
1	@Gere341	TAKE NOTE LIBS- **Canada** is a gun controlled country. Yet there was a deadly Mall Shooting last June Someone wants 2 get access2guns, theyWill	16
1	@RightCentrist	466 violent crimes per 100K ppl in the US, 2034 violent crimes per 100K ppl in **UK** - Statistic says it all - gun control fails.	14
2	@alexblagg	22 children in **China** attacked with a knife today, no deaths. Senseless violence can't be prevented. Gun violence can.	182
2	@jasonwstein	18 school shootings in **Australia** before 1981. They banned semi-automatic weapons. No big school shootings since. via @cnbc	19
2	@igorvolsky	9,000 people killed with guns last year. In similar countries like Germany, 170. **Canada**, 150. There is a reason for that.	12
2	@Tinkerbell_	51 people were killed by guns in the **UK** in 2011. 51. In the ENTIRE year. USA 8,583. Now say gun control does not work.	10
3	@EstherKramer1	22 hurt by knife attack in **China** vs. 20 kids dead via gun in USA. Crazy people kill but guns help them out a lot. #GunControlNow	14
3	@rationalists	1996: Gunman kills 35 in Port Arthur, **Australia**. 1997: Australia bans guns. 2012: Massacres since 1997: NONE!	57
3	@sean_dixon	In **Canada** we watch the same films as in the US. We play the same games, have the same mental health issues. So whats the difference? #NRA	11
3	@Good_Beard	5 times as many murders per head of population in USA as in **UK**. Do you think thats because Americans are 5 times as evil or have more guns?	53

a clear pro-gun rationale, and an anti-gun position for groups 2 and 3 is also evident. The content posted by users, therefore, is in accordance to the measurement of our polarization metric. More interestingly, each group attempted to use statistics from each country in their favor; the same country is used as a case that favors gun rights (group 1) and as a case that favors gun control (groups 2 and 3). The focus on evidences that reinforce previous opinions is a cognitive bias known as confirmation bias [127]. In the case of China, Twitter users comment on the *same* fact – the attack of children with a knife-armed man – and yet they use the fact to reinforce contrasting opinions. Such phenomenon, known in the social psychology literature as *belief polarization* [109], is one of the strongest evidences that a group of individuals is divided into polarized groups. Note that our understanding of the relationship between the three groups provided by our polarization metric P allowed us to quickly find such contradicting opinions, and our methodology may support sociological studies on polarization of opinions based on social media data.

2.4 Concentration of Popular Nodes Along the Boundary

In this section we investigate another structural characteristic that may help on the identification of polarization – the concentration of popular (high-degree) nodes in the boundary.

Since polarization is associated with antagonism, we expect popular nodes to be present far from the boundary, as strong representatives of their group viewpoints that do not find endorsement from the opposing side. On the other hand, we expect non-polarized communities to promote the existence of high-degree nodes in the boundaries, since such nodes are more prone to enjoy popularity from both sides.

To measure the concentration of high-degree nodes in the boundary, we build, for each social network, two ranks r and r_b. r is a rank of all nodes in the graph sorted by degree, in descending order of popularity. r_b ranks the same nodes, but according to d_b, i.e., the number of cross-boundary connections. We then use Spearman's rank correlation coefficient [147] ρ to capture the statistical dependence between r and r_b. Spearman's correlation captures how well the relationship between two variables can be described by a monotonic function and its value ranges from -1 to $+1$. $\rho(X, Y) = 1$ means that variable Y is a perfectly monotonic function of X. In our context, a high ρ means that high-ranked nodes in the graph tend to be also high-ranked in the boundary, indicating a concentration of high-degree nodes along the boundary. A low ρ indicates that many high-degree nodes in the graph are low-ranked in r_b, what indicates that there is a significant number of popular nodes which do not belong to the boundary.

Figure 2.6 compares overall and boundary ranking positions for nodes in the `University` social network. Note that high-ranked nodes in r tend to also be high-ranked in r_b, and $\rho = 0.84$ indicates that the network promotes a convergence of popular nodes to the boundary. We interpret this result as a strong indication of absence of polarization.

In Figure 2.7 we show r and r_b for the nodes belonging to non-polarized communities 2–3 in the gun control network. This graph is better interpreted when compared to Figures 2.8 and 2.9, which exihibit the corresponding results for polarized communities 1–2 and 1–3, respectively. Since nodes that exhibit the same degree are tied in the rankings, we added to each rank position a random value between 0 and 5% of its absolute value to allow a better visualization of point density. Note that, in Figure 2.7, a large number of high-ranked nodes in r are also high-ranked boundary nodes in r_b. A large concentration of nodes is observed in the range 1–5000 of r and r_b in this pair of communities, in comparison to Figures 2.8 and 2.9. The ρ value is also significantly higher in the case of Figure 2.7 ($\rho = 0.70$), supporting our intution w.r.t. the relationship between the concentration of high-degree nodes along the boundary and the existence of polarization.

Table 2.5 shows ρ measurements for the other social networks we consider in this work. We note that, although polarized networks tend to exhibit lower values of ρ, this is not always true. The U.S. Political blogs has a concentration of popular nodes in the boundary which is equivalent to the NYC-Teams network, despite of the differences in both Q and P. A possible explanation for such differences is that the political domains count with many

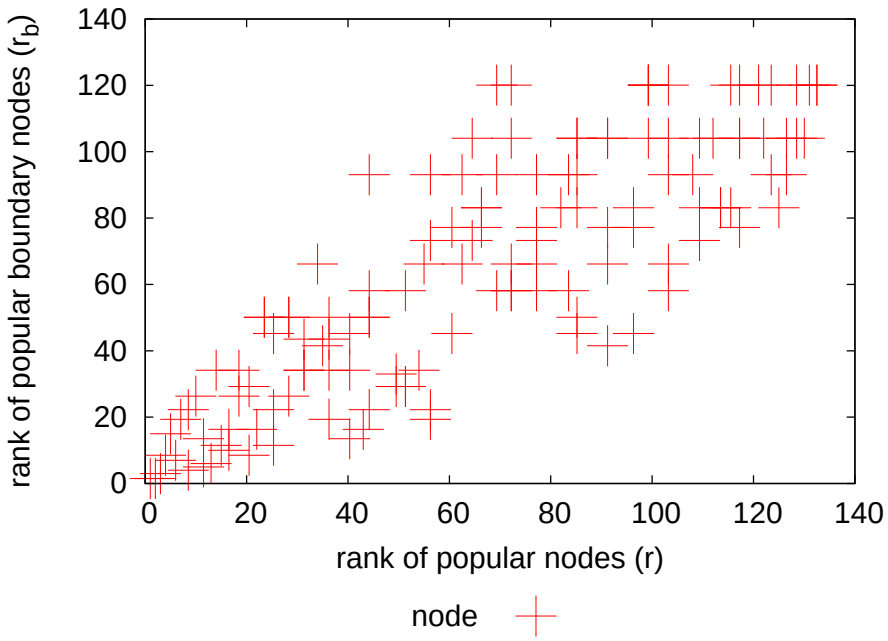


Figura 2.6. Concentration of popular nodes in the boundary for `University` network – $\rho = 0.84$. The high value of ρ indicates absence of polarization.

Tabela 2.5. Modularity Q, Polarization P and Spearman's Correlation ρ for networks described in Table 2.1.

network	**media**	**modularity** Q	**polarization** P	ρ
1 - NYC Teams	Twitter	0.15	-0.002	0.65
2 - University	Facebook	0.24	-0.24	0.84
3 - Karate's Club	friendships	0.35	0.17	0.62
4 - Brazilian Soccer Teams	Twitter	0.39	0.20	0.39
5 - US Political Blogs	blogs	0.42	0.18	0.65

media outlets that connect to both sides and thus gain popularity in the boundary, despite the polarized context.

2.5 Conclusions

In this part of this work, our goal is to demonstrate that literature of polarization of opinions in social networks has focused attention on domains *previously known* to induce polarization; as a consequence, the necessary and sufficient structural characteristics of polarized social networks were unclear. We perform a comparison between polarized and non-polarized networks and propose a new metric designed to measure the degree of polarization between two communities. Unlike modularity, which simultaneously measures homophily and antagonism between groups, our metric focus on the existence (or absence) of antagonism

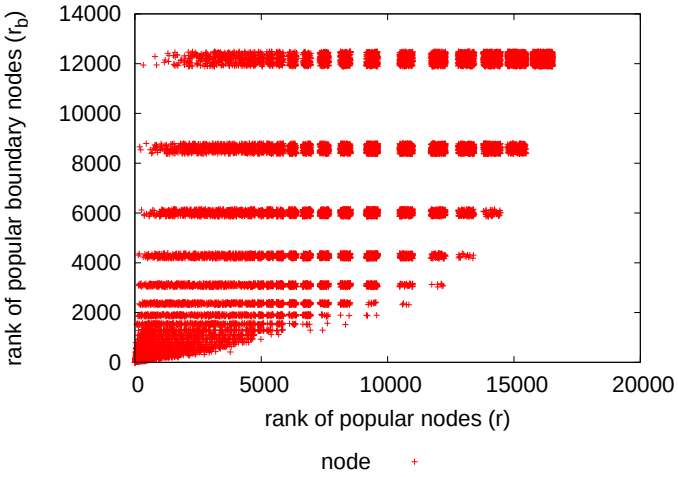


Figura 2.7. Concentration of popular nodes on the boundary – Gun Control communities 2–3 – $\rho = 0.70$. The high value of ρ indicates absence of polarization.

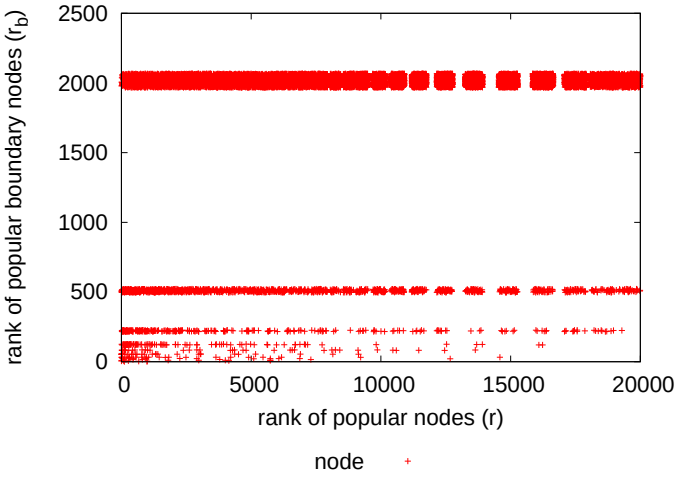


Figura 2.8. Concentration of popular nodes on the boundary – Gun Control communities 1–2 – $\rho = 0.21$. The low value of ρ indicates presence of polarization.

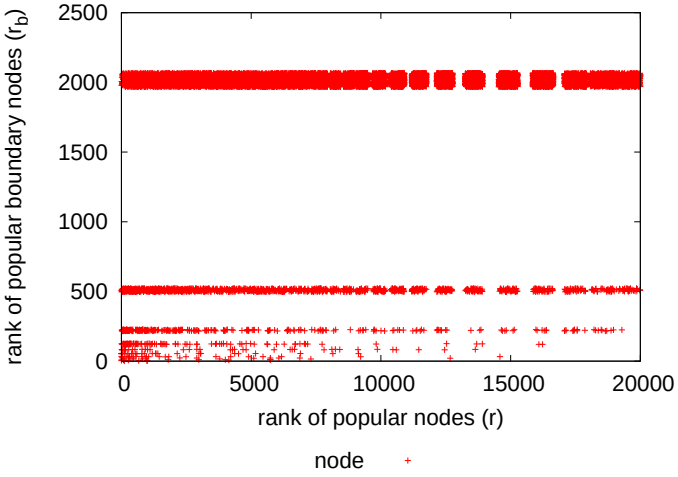


Figura 2.9. Concentration of popular nodes on the boundary – Gun Control communites 1–3 – $\rho = 0.23$. The low value of ρ indicates presence of polarization.

between the groups. We consider nodes' decisions towards connecting to users who belong to the other (potentially opposing) group, in comparison to connect to members of its own group. Furthermore, we also show that polarized networks tend to exhibit a low concentration of high-degree nodes in the boundary between two communities.

One aspect that we do not took into account in our metrics is the group sizes; they make the implicit assumption that group sizes to not differ significantly. Extending the formulas to account for group sizes is left to future work.

In practical applications, we believe that modularity and our metrics P and ρ can be used together and complementarily, helping in raising evidence whether a topic discussion embedded in a social network is subject to polarization. As we will demonstrate in the next Chapters, the particular structural characteristics of a polarized social network enable tasks such as sentiment analysis to be addressed based on well-known behavioral patterns that arise in such a setting.

Capítulo 3

Finding Community Structure and Community Relationships in Multipolarized Social Networks

As we discussed in Chapters 1 and 2, a polarized discussion is often related to a strong bias on opinion holders. Bias has been extensively studied by sociologists as a characteristic of people or entities which holds a *partial* view regarding a topic or issue and therefore lacks neutrality. We can find bias in almost any scenario where opinions are expressed, and common contexts where bias can be easily spotted are Politics, Sports and debate on public policies [84]. It is observed both on ordinary people and on the mainstream media [119], which can be biased both when selecting which events are covered and on the veracity of their coverage [45].

In general, biased opinion holders exhibit one or more of the following characteristics [161]:

- lack of proper balance and neutrality in argumentation;
- lack of proper critical doubt;
- the existence of a personal interest from the arguer in the outcome of the argument or discussion.

Taking decisions based on biased judgments is a pervasive characteristic of human behavior which results from *cognitive biases*, which are systematic errors on judgment and interpretations [13]. Such errors can be produced by emotional and moral reasons (such as the case when the arguer has an interest on the issue), but also can be caused by subtle, irrational characteristics of human reasoning. A wide list of cognitive biases that affect human thinking and how they express their opinions about facts have been identified in the

last six decades by social scientists and psychologists [134]. Some of the most pervasive biases that affect people's opinions are:

- **Confirmation Bias:** Confirmation Bias is a cognitive bias which makes people interpret new evidence in ways that confirm and reinforce the beliefs they *already* hold [114; 133]. It is sometimes referred as one of the main flaws on human reasoning [127]. It causes a selective thinking whereby one tends to notice and to look for whatever confirms his or her beliefs, and to ignore, not look for, or undervalue the relevance of what contradicts these beliefs. A classical example of confirmation bias on a discussion is the case of soccer fans which are sure that their teams are always in disadvantage in referees' decisions: since they already hold this (negative) opinion, there is a high chance they will pay attention on subsequent referee's mistakes against their teams and do not pay the same attention to mistakes in favor their teams.

  Another interesting example on how confirmation bias manifests was discovered by researcher Valdis Krebs while analyzing book purchasing trends during the 2008 U.S. Presidential Elections. He found that people who already supported Barack Obama tended to buy books that praised him, while people that disliked Obama had an increased chance of buyings books which were critical to him [86]. In other words, if someone has a favorite presidential candidate, he or she is more likely to pay more attention and give more credence to information that is favorable to him, and negative to other candidates [83], and such confirmation bias will inevitably lead us to give more positive opinions on people and subjects we already support and more negative opinions on what we already were against.

- **Motivated Reasoning:** Motivated reasoning is the process of subjecting information that contradicts our previous beliefs to greater scrutiny than information that confirms our existing beliefs [162]. Motivated reasoning is a clear manifestation of the fact that the goals and motives one have in mind affect their reasoning [87]; in other words, we prefer to believe in our existing beliefs because it is "easier" than examine new contradictory information, in an effort to maintain existing evaluations. Reasoning away contradictions is psychologically easier than revising our feelings. In Politics, motivated reasoning has been shown to manifest when voters acquire information and determine whether that new information supports or opposes their candidate expectations. Instead of decreasing their good evaluation on their preferred candidate, the opposite may happen [136].

- **Hindsight Bias:** Hindsight bias is a cognitive bias that is related to the tendency people have to view events as more predictable than they really are. After an event has taken place, people often believe that they knew the outcome of the event before it

happened [114; 162]. Hindsight bias is also known as the "I-knew-it-all-along phenomenon". This bias affects opinions since people criticize other people and facts **after** the event has happened, suggesting that they had a prediction capability that they actually do not have. Such bias is pervasive in domains such as Sports, in which opinions such as "I knew the Giants would defeat the Patriots in the Super Bowl" and "I knew they should have changed quarterbacks" are commonplace [152].

- **Self-Reporting Bias:** When people report to others about their feelings and the daily events that happen with them, they are often biased in the sense that what they report is not a random sample of what they feel. This is specially true in social media systems, where people make a decision in posting an opinion or not depending on whether they think it will be of interest of their friends [82]. An example of such self-reporting bias is that we are more likely to see people tweeting about drinking wine than drinking water, even though the latter action is much more frequent than the first. For the matter of analyzing opinions, we have to keep in mind that people can be more motivated to express positive or negative opinions depending on the scenario, and this bias can affect the interpretation of sentiment expressed in social media systems [78].

Note that, in different ways, most cognitive biases lead the opinion holder to keep, justify and reinforce a *previous* opinion or belief, while new events happen regarding the issue he or she is judging. Sometimes bias is so strong among opinion holders that it can lead to two people with different biases draw different conclusions after examining the *same* evidence, a behavior known as *attitude (or belief) polarization* [8]. In a seminal experiment conducted by [109], subjects who were selected because of having different views on the death penalty were pulled further apart after reading the same essay about the death penalty.

The connection between cognitive biases and polarization is that, on a polarized debate, most people already have such "previous" opinion: it is their favorite candidate/party, their preferred soccer team, if they are in favor or against some governamental decision etc. In some cases of extreme bias, the opinion holder can be considered almost as a proxy for the opinion itself [150]. For instance, someone who clearly supports a candidate in an election will tend to post positive comments about him and negative comments about his/her adversaries on a regular basis.

As a consequence, quantifying the opinion holder bias towards a topic may be of great help in predicting opinions. On the remainder of this Chapter, we assume that opinion holders sharing a similar viewpoint will cluster themselves into communities in a social network, and devise a method to accurately find the communities themselves and the polarity relationships among communities – a crucial information on predicting opinions regarding entities that belong to individual communities, as we will later demonstrate.

3.1 Finding communities and community relationships in social networks

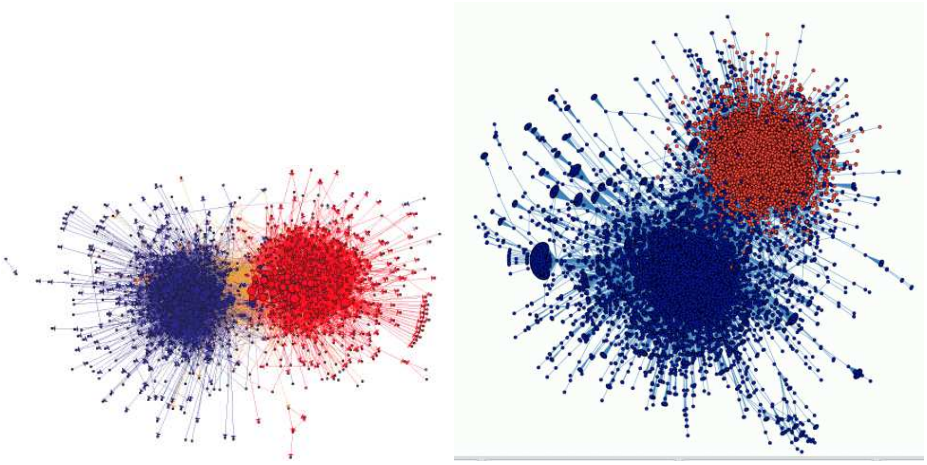
Directly identifying the manifestation of cognitive biases listed on the previous section would be challenging, since each cognitive bias would need to be modeled and captured separately. Our strategy is to make use of the *social interactions* among social media users to try to infer their individual biases regarding a polarized topic T. We hypothesize that opinion holders carrying a similar viewpoint with respect to T will naturally group themselves into communities – a set of clusters containing nodes that establish links more frequently to nodes belonging to the same cluster than with nodes from different clusters [47]. The study of algorithms that find meaningful communities in graphs and more specifically in social networks is an established and field in network analysis; most of these algorithms seek to maximize a criteria of link density inside communities and minimize it across communities, leading to well-known approaches such as modularity maximization [59], random walks and spectral analysis [171].

Standard community mining on unsigned graphs, however, do not output the relationship among the K communities found. Is the relationship between each of the $\binom{K}{2}$ pairs of communities antagonistic, supportive or indifferent one to each other? Finding such relationships is of great importance not only to the social sciences, but also to support the design of algorithms that exploit the network structure in conjunction with opinionated text expressed to better perform tasks such as recommendation, sentiment analysis and news curation on online social platforms [29; 150; 53; 103], as we will later detail in this dissertation, in Chapter 4. In the social network analysis literature, community relationships are usually found through different approaches:

- **Inferred from the domain.** In many contexts, it is previously known that the domain of discussion induces polarization among a pair of communities and stimulate a social group to divide itself into two sub-groups with conflicting viewpoints regarding a topic. Political ideology, same-sex marriage, gun control, abortion, immigration and global warming are only a few examples of scientific, moral and social divisive issues that are known to become dominated by increasingly extreme and strong opposite opinions [92; 121; 73; 163] and generate well-separated social communities, as shown in a diverse set of social media systems such as Twitter and the Blogosphere [3; 33]. As an example, Figure 3.1 shows the division on opinion holders writing about U.S. Politics on two different social media platforms. Figure 3.1(a) (extracted from [3]) displays the U.S. political blogosphere during U.S. 2004 Elections, while Figure 3.1(b) shows a similar pattern on Twitter, during the 2010 U.S. Congressional midterm Elections ([33]). In

such settings, no specific analysis on the polarity of the links crossing the communities is performed; the antagonism is implicitly assumed due to the political domain and the modular division of the social graphs into two communities historically known to be antagonistic.

- **Inferred from a signed graph.** Uncovering polarity relationships among communities do not bring significant additional challenges apart from finding the communities themselves if we have access to a *signed graph* – a network whose edges are explicitly labeled as positive or negative [97; 98]. If that is the case, communities can be found through algorithms designed to mine signed networks such as [91; 170]. After the communities are determined, the proportion and volume of positive and negative edges flowing between each pair of communities will directly unveil the relationships of support, antagonism or indifference between communities [171].



(a) 2004 U.S. Political Blogosphere (b) 2010 U.S. Political Twitter

Figura 3.1. On the left, a typical bipolarized social network showing the division of political blogs into two communities – liberals and conservatives [3] (edges are web page citations). On the right, a similar division on Twitter (edges are retweets). The separation of nodes into two clear communities is a strong pattern observed on polarized discussions. Usually, no explicit analysis on edge signs is performed and the antagonism is assumed because of the domain of analysis.

In this Chapter we aim to analyze social networks which do not fall in those two cases. We are particularly interested in inferring polarity relationships among social communities in settings that follow the following properties:

1. **Multipolarization ($K > 2$).** While the majority of research on social network analysis of polarized discussions has focused on the classical case of bipolarization – characterized by the emergence of exactly two conflicting groups representing two opposite

viewpoints, as shown in Figure 3.1, we are interested on domains where more than two viewpoints and discussion sides that interact with respect to a topic arise: think, for example, of multipartisan political systems (as in Brazil). In Figure 3.2, we plot in different colors the three largest communities found in a network of retweets and replies obtained from Twitter during the 2014 Brazilian Presidential Elections, representing groups of people formed around the 3 main candidates (Dilma Rousseff, Aecio Neves and Marina Silva). Differently from the bipolarized political landscape from Figure 3.1, on multipolarized discussions we observe more complex relationships and interactions between social groups, rather than the support versus antagonism dualism. On the bipolarized case, once you find the leaning or preference of a user or group toward a topic or issue, their viewpoint regarding the opposite viewpoint is implicitly determined. For instance, the determination of the community of supporters of pro-choice in abortion discussion implicitly carries their antagonism and disagreement with respect to the pro-life side; the same rationale applies to the division of nodes into democrat/republicans, pro gun-control/pro gun freedom etc. However, when there are $K > 2$ possible sides one can belong to, the identification of an individual as a member of a community does not necessarily implies on a notion of antagonism with respect to all the remaining $K-1$ groups. He/she can be indifferent, or neutral, to a subset of the remaining groups, or can support more than one group simultaneously. On a multipolarized social graph, once we find that an opinion holder belongs to a community, his or her preference regarding the other sides is not automatically determined and we need to explicitly find the negative opinions regarding each side, if they exist. In this Chapter, we propose an automatic method of finding negative messages posted in a social network formed by (potentially) antagonistic communities which requires as supervision a small set of positive seeds (i.e., users or messages) that convey a positive polarity with respect to each community.

2. **Unsigned network.** On general purpose online social networks such as Facebook and Twitter, there is no explicit positive and negative signs encoded on the edges, and for most tasks all edges are assumed to be positive [173; 88]. Edges in such networks are created by two main types: message broadcasts (e.g., retweets, shares) and communication interactions (replies, or comments). Both are inherently ambiguous regarding to the polarity of the sentiment they convey. Replies, as on web hyperlinks, do not carry an explicit sentiment label and can be either positive or negative [97; 176]. A message broadcast, on the other hand, tends in most cases to be a signal of agreement; in fact, first works on behavioral analysis on Twitter defined retweets as a strictly positive interaction [26]. It is known from empirical observation, however, that people also use

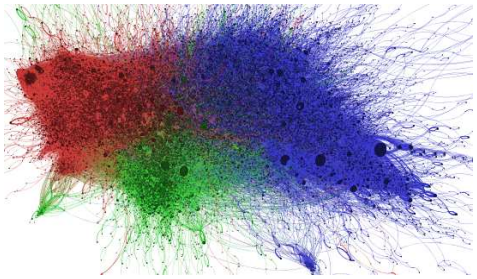


Figura 3.2. Network of retweets and replies obtained from Twitter show 3 communities formed around the 3 main candidates in the 2014 Brazilian Presidential Elections. The $\binom{3}{23}$ pairs of communities do not necessarily share mutually antagonism, and antagonism for each pair can exist in different intensities. Comparing different levels of antagonism between each pair of community is a challenge that is not present on bipolarized social networks.

broadcasts to convey a negative sentiment with respect to the message's author or its content. "Retweets are not endorsements" is a common line found in biographies of journalists and think tankers in Twitter, while some people share stuff that they vehemently disagree with only to show the idiocy of the people they oppose. In summary, retweets and shares can be used as a "hate-linking" strategy – linking to disagree and criticize, often in an ironic and sarcastic manner, rather than endorse [154]. One can also broadcast the original message and add comments to it, often in disagreement with the original content, what also contributes to turn retweets into an ambiguous signal with respect to the sentiment being conveyed in that interaction.

In the context of multipolarized communities in unsigned networks, we make two main contributions. First, we demonstrate that the simplifying assumptions valid for bipolarized social networks do not hold when we go to the $K > 2$ case. Multipolarized social networks unveil subtleties and inconsistencies that are "hidden" on traditional bipolarized networks on which current research focuses. In particular, we found that communities that are **more** antagonistic share each other's content **more often**, what can be easily misinterpreted as a signal of support by naïve network models. On a bipolarized social network, such behavior does not manifest as a problem, but as we will show later in this Chapter, it significantly harms the understanding of group relationships when more than 2 communities arise with respect to a topic.

Our second contribution is to propose a strategy to make sense of group relationships in multipolarized social networks, in settings where the two aforementioned approaches are not applicable, i.e., the domain does not imply that antagonism is the dominant relationship between every pair of communities, and edges signs are not natively available. To make sense of the relationships between multipolarized communities and deal with the ambiguity

in the signal provided by retweets and replies, we propose to employ a more reliable signal to detect antagonism: the *lack* of interaction between some specific sets of users and messages. These sets are defined in a way that only messages that lie on the boundary between communities are considered, in order to provide a stronger confidence that users that do not react to a message are doing so because of disagreement, not because they are not aware of such messages. Such messages, inferred from a negative implicit feedback strategy, are then considered as *negative seeds*, from which we propagate random walks that identify the degree of antagonism with respect to an entity in the whole social graph.

This Chapter is organized as follows: Section 3.2 discusses related work on polarization, signed networks and analysis of antagonistic communities. Section 3.3 analyzes two Twitter social graphs build over interactions on Politics and Soccer topics to empirically demonstrate that, on multipolarized social networks, the naïve assumptions made by analysis on bipolarized social networks are ambiguous and misleading. On Section 3.4, we present our model that correctly identifies the relationships among multipolarized communities.

3.2 Related Work

Our work focuses on community detection and the relationships among the communities in a setting that lies in the intersection of two fields in the social network literature: social networks subject to polarization and signed networks.

From the sociological perspective, polarization can be formally understood as a state that "refers to the extent to which opinions on an issue are opposed in relation to some theoretical maximum", and, as a process, it is the increase in such opposition over time, causing a social group to divide itself into two sub-groups with conflicting and antagonistic viewpoints regarding a topic [149; 74; 121; 132]. Understanding polarization on online discussions and the social structures induced by polarized debate is important because polarization of opinions induces segregation in the society, causing people with different viewpoints to become isolated in islands where everyone thinks like them [159]. Such filter bubble caused by social media systems limits the exposure of users to ideologically diverse content, and is a growing concern [95; 11].

Polarization has been measured when interactions are known to have a predominant positive or agreement tendency; for instance, one can measure media bias by simply counting the number of times a particular media outlet cites various sources, and compare this to the citation rates of those same sources by congressmen; this approach, for instance, unveiled a strong liberal bias in the US news media [119]. In such analysis, there is a strong assumption, not always made explicit, that a citation is on average positive. Studies adopting similar

assumptions have categorized political blogs according to their political bias and analyze the communication patterns and network structure that different political views induce [3]. Bias in political blogs has also been used to predict the bias of political articles in the online news media [53], by, again, counting the number of liberal and conservative blogs that cite each article. A similar strategy has been employed by [61], but considering Twitter followers as the primary evidence of bias.

The vast amount of work on polarized social networks, both on the social and computer sciences, limit themselves to analyze the traditional case of two conflicting sides: liberal versus conservative parties, pro-gun and anti-gun voices, pro-choice and pro-life [33; 107; 3; 175; 169]. Our work aims to characterize and model scenarios where more than two communities respond and discuss a topic, as a way of clarifying these hidden assumptions made by polarized social networks analysis that have been conducted on previous research, as will become clearer on the next sections.

Our research also relates to another specialization of social network analysis which so far has been done independently from polarization studies – signed graphs. Structural network characteristics such as distances, clustering coefficients and centrality have been measured in signed social networks [88] and supervised approaches that aim to classify edges into positive or negative have been of great interest recently [98; 97]. When the edge signs are known, extending community detection algorithms to deal with negative edges is a natural path that has received some attention recently [91; 170; 171; 108]. In such cases, the relationships among communities is easily reflected by the number of positive and negative edges flowing from the source community to a target community.

Since we are interested in social networks extracted from platforms such as Twitter and Facebook, where broadcasts and replies can be either positive or negative, community detection algorithms that receive as an input a signed network are not directly applicable. Very few works try to infer edge signs from a network of unsigned edges [173]; some works explore evidences of agreement or disagreement to predict the sign of the edges in the absence of ground truth information. For example, [2; 9] build NLP models based on the surrounding text of a paper citation to predict the polarity of a citation, while [158] have explored edits on wikipedia content as evidence of disagreement between users. [4] have noticed the fact that comments and replies on newsgroups tend to indicate disagreement, a consequence of the fact that when you disagree you have more to say then when you agree; agreement usually implies on being redundant on what has already been said [63]. Another line of research has been to mine the text [66; 67] associated with users communication to infer their relationships.

While many works recognize that negative links are generally not explicit in social media, many works try to infer them assuming that the observed interactions are posi-

tive [89; 151]. While retweets are still widely seen as a positive interaction, more recent works started to investigate negative aspects of retweeting, such as fake retweets that artificially boost users' popularity [57]. Our analysis, however, see retweets with a negative polarity as legitimate interactions. Many users, indeed, stress in their bios that "retweets are not endorsements".

Our work borrows from both lines of research in social network analysis because, although we still base our work to domains subject to polarization that lead to the formation of communities, although we do not automatically assign antagonism among all pairs of communities. On the other hand, we do not have access to a signed graph, we do make assumptions regarding the distribution of edge signs – namely, that users, most of the time, establish interactions that are positive with respect to their point of view.

3.3 Community mining on a network of retweets

Our goal in this Section is to empirically demonstrate what are the implicit assumptions assumed by network analysis of bipolarized unsigned social networks and how they lead to misleading conclusions when applied to networks that emerge from online discussions that induce the division of opinion holders into more than two communities.

We used Twitter's Firehose API to monitor a topic that motivate intense debate and discussion on online media in general and thus are suitable for analysis of formation of antagonistic and conflicting communities: Sports [93; 166]. More specifically, we collected tweets about the 2010, 2011, 2012, 2013 and 2014 editions of the Brazilian Soccer League. We collected mentions to the 12 largest Brazilian soccer teams and related keyword, such as goal, penalty, yellow card, offside, among others. Table 3.1 provides details on the dataset.

Tabela 3.1. General description of the dataset on Brazilian Soccer debate collected from Twitter.

information	**Brazilian Soccer dataset**
period	2010-2014
# entities	12
# *tweets*	107.0 million
# *retweets* (RTs)	22.2 million
# *replies*	8.3 million
# users	14.1 million

Different graphs may be built based on these data. We chose to run the methodology that is more commonly adopted by the literature: a graph $G(V, E)$ is built where V is the set of users and E is the set of directed edges, where (u_1, u_2) is in E if u_1 has retweeted or replied to a message posted by u_2 during the period of analysis [3; 33]. For this graph, we

run a standard community detection algoritm (MCL [24]) to find groups of users that interact more frequently within the group than with members from other groups. As expected, users self-organize into communities around the most relevant entities associated to each topic. Figure 3.3 displays the communities formed around the largest Brazilian soccer teams.

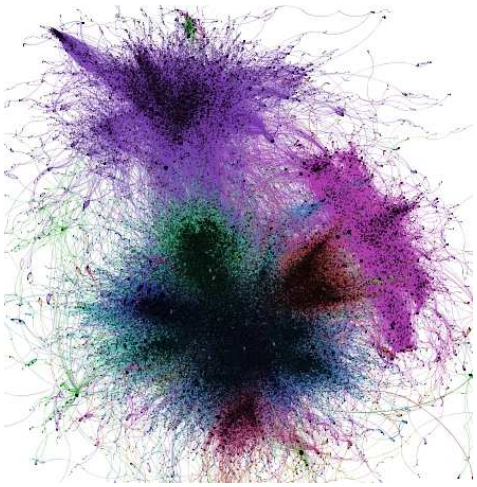


Figura 3.3. Network of retweets and obtained from Twitter showing communities formed around each of the top Brazilian Soccer teams.

Once the communities are found, our goal is to understand the polarity of the relationships among each pair of groups. Recall that, on polarized domains where two communities are found, no subsequent analysis is usually performed, other than the quantification of the degree of separation between the pair of communities, using community quality metrics such as *modularity*, as we pointed out in Chapter 2. It is a standard practice to assume that the more separated the communities are, the more antagonism is observed, as a consequence of the homophily principle [113]. For instance, [181] correlates the increase in modularity in the network of congressmen in the United States over time with the increase of polarization between Republicans and Democrats. Those studies are constrained because there is only one value of modularity or number of edges flowing from one community, hindering the capability of deeper understanding the semantic or value of such a number. Since the only pair of communities from the domain to have an antagonistic relationship is already known, it is not clear how the distribution of interaction types and the intensity of the interactions relate to the polarity of the group relationship.

However, since we work with $K > 2$ communities, we now have $\binom{K}{2}$ pairwise community separation measurements to compute and compare. More specifically, we compare the proportion of retweets triggered from users belonging to community i that flow toward community j:

$$prop(i,j) = \frac{RT_{i,j}}{\sum_{k=1}^{K} RT_{i,k}} \tag{3.1}$$

To evaluate the group relationships based on ratio of retweets flowing between each pair of communities, we use as "ground truth" the known local rivalries that exist in Brazilian Soccer among soccer clubs from the same city, as listed on Table 3.2. We do expect, thus, that the relationships among the communities which concentrate supporters of each team reflect, in some sense, the stronger antagonism known to exist due to local rivalries.

Tabela 3.2. Local Rivalries in Brazilian Soccer. Stronger antagonism exists between soccer clubs and supporters belonging to the same Brazilian state.

state	**local rivalries**
Minas Gerais	Cruzeiro, Atletico-MG
Sao Paulo	SPFC, Santos, Corinthians, Palmeiras
Rio Grande do Sul	Grêmio, Internacional
Rio de Janeiro	Flamengo, Fluminense, Vasco, Botafogo

On Figure 3.4 we plot the distribution of of $prop(i,j)$ for all the $\binom{K}{2}$ pairs of communities formed around supportes of Brazilian clubs. The stacked histograms show an unexpected result: communities that are more antagonistic one to the other can broadcast each other's content *more often* than when there is less, or none antagonism between them. For example, the community of users that Cruzeiro supporters more frequently retweets is Atlético-MG, their fierce rival in Minas Gerais state. In Rio Grande do Sul, Internacional supporters also retweet Gremio's tweets very often, even through they are also fierce rivals.

Note that, on traditional bipolarized domains in which current literature focuses, such inconsistency is not noticed at all, since there is only a single pair of antagonistic communities and thus only a single separation/interation metric to be computed.

The empirical observation that antagonistic communities share each other's content more often than expected can be explained by the negative semantic that can be embedded in retweets and broadcasts, such as:

1. **Sarcasm.** It is common to see a user propagating a message he or she disagrees with and putting it out of context, in order to create sarcasm or irony. In this case, we usually see messages shared a certain period of time after it was originally posted. A classical case is when the original message made a prediction that turned out to be shown false some time later.
2. **Fake or edited retweets.** Another practice which is common among Twitter users with contrary views is to create fake retweets, in the format "*RT @user fake message*",

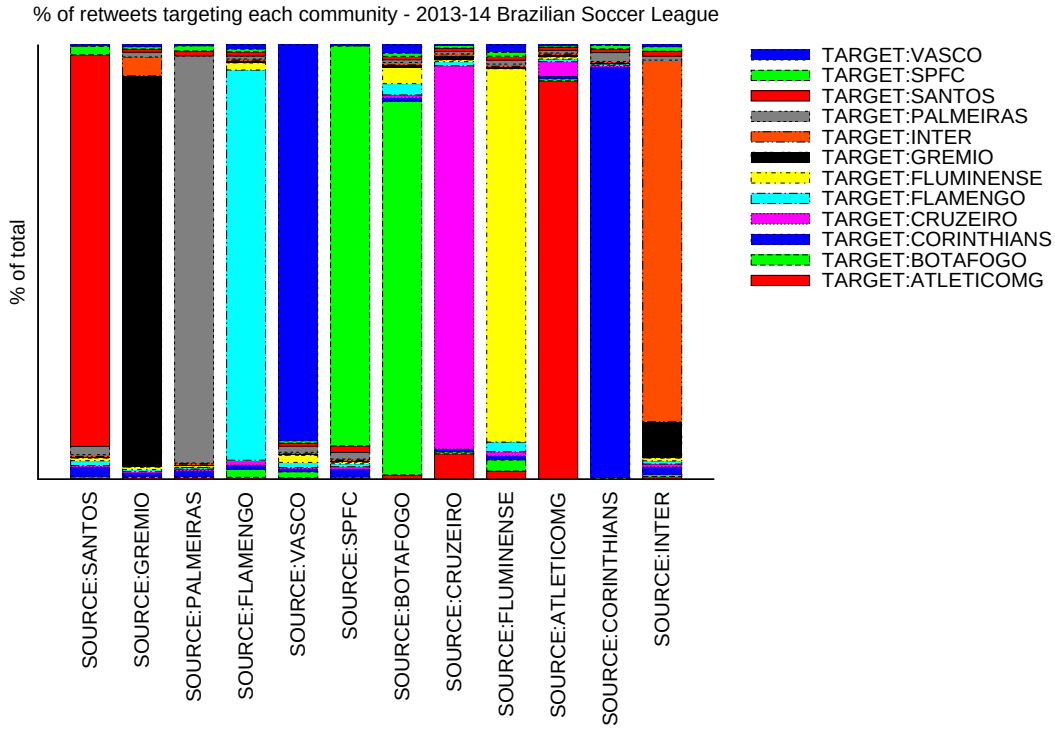


Figura 3.4. More antagonistic communities retweet each other more than indifferent communities (see stacked histograms in conjunction with Table 3.2. Sources and target of retweets are shown.

assigning to *@user* a message he/she never has posted. Fake retweets have already being investigated as a spamming activity in Twitter [122], in which spammers try to borrow from the reputation of celebrities. In the context of polarized discussions, however, the goal is different – to create humour or to make criticism, or even spread false information [124].

3. **Share to show contrary opinion.** Many times, a user propagates a message he or she disagrees with to show the message to his/her followers or friends and comment on that content. The goal is to start a discussion and gauge reactions.

The ambiguity of message broadcasts and replies can cause nodes to be wrongly assigned to its community, but, in general, it is not a big issue in finding communities because on average these interactions are positive, and thus link density-based approaches are able to correctly find the groups.

In this Section, we highlighted two assumptions that usually are implicitly made in community analysis on unsigned social networks, which we here make explicit:

1. It is implicitly assumed that interactions are, on average, more likely to be positive than negative, what leads to the correct identification of groups, despite the fact that some of

the interactions may be, indeed, negative. For instance, in Figure 3.1(a), one blog can cite the other to disagree with it, but since most of the time blog citation is used in a positive way, the two groups are found. Since a fraction of retweet edges are negative, it can cause misclassification of some users with respect to the community they belong to.

2. The presence of antagonism is usually implicitly assumed from the domain, rather than inferred from a principled method. Once users are grouped into two communities, members of one group will automatically be assigned to have a contrary or antagonistic opinion regarding the remaining group. These works do not need to deal with differences between antagonism or indifference, neither with a more accurate treatment of edge signs.

3.4 Semi-supervised community detection

In this Section, we use the empirical observation learned on the previous section to devise a community mining algorithm that outputs a set of K communities and the relationships among them. We work in the context of a well-delimited topic T where a certain level of antagonism is expected among a subset of the communities induced by T. Moreover, we assume that the number of communities K is known in advance and it is a parameter of our method. For instance, a typical topic of interest is if T="2012 US Elections", $K = 2$ and $\{K_1 = Democrats, K_2 = Republicans\}$ are the possible sides one can belong to. In case of T="Brazilian Soccer", $K = 12$ if we take into consideration communities that support the 12 most relevant Brazilian soccer clubs.

We want to exploit social interactions established in the context of T to solve the following learning problem:

Given: K sides of discussion, a set of users U, a set of messages $M = [m_1^{u1}, m_2^{u2}, ..., m_n^{un}]$, where u_i is the author of the message (which belongs to the U set), and a set of relationships $E \subset U \times M$ that induce a bipartite graph G. These relationships can be of multiple types, e..g, a user broadcasted a message (retweet, share) or replied to a message. No individual sign of edges is provided.

Estimate: $P_{ui}, \forall u \in U, 1 \leq i \leq K$, and $P_{mi}, \forall m \in M, 1 \leq i \leq K$. P_{ui} and P_{mi} represent the probability that user u or message m provide a positive view with respect to the viewpoints represented by community i.

$P = [P_u P_m]^T$ is a matrix that, for each pair (e, k), where e is a node in G (which can be either a user u or a message m), k is a community, quantifies at $P_{e,k}$ the probability that e leans towards community k. Note that this representation naturally allows users and

messages to have soft memberships to each community. For instance, if T="US Elections", a given media outlet m can be found to be $P_{m,democrats}$ = 0.40 and $P_{m,republicans}$ = 0.60, indicating that it tends to be fairly negative with respect to democrats (0.40 < 0.50 – which represents full neutrality), and fairly positive to republicans. This pattern would differ, for instance, from the official profile of Hillary Clinton, which could have $P_{m,democrats}$ = 0.99 and $P_{m,republicans}$ = 0.10 as reasonable values.

Notice that, at the same time that we seek to learn the leanings of each user/message with respect to each $k \in 1 \leq i \leq K$, communities are naturally found in this framework by assigning node e to community k according to $argmax_k = P_{uk}$, i.e., the community to which the user is more positively leaned.

Our modeling has three main features:

Soft membership. There is a recent trend in the community detection literature to focus on *overlapping communities* [172; 165; 129], which allow a node in the graph to belong to more than one community. Grouping a node into a single community is a too strict decision for many nodes; moreover, one node (either a user or a message) can support more than one discussion side.

G as a bipartite graph of users and messages. It is known that the definition of the network that will represent a set of data can greatly depend on the kind of task one is trying to solve [37], and different networks relating users and their interactions could be built based on the stream of messages we collect. Traditionally, a social network $G(V, E)$ is represented as a set of users V and a set of edges E connect two users if they have interacted at least once; thresholds can be applied to filter less frequent interactions [37]. The problem with this network is that it hides user-message interactions that happen in the network: for instance, two users with opposite opinions may propagate different messages from the same media outlet, what could wrongly indicate that both share the same opinion. Connecting users in this way hides the fact that each user post messages with a potentially different sentiment with respect to different entities; i.e., a media outlet can post a positive message to the republican candidate one day and a negative message a week later. We then choose to represent interactions (retweets, replies) in a user-message bipartite graph, as shown in Figure 3.5. In this graph, the set of nodes V is composed of two disjoint sets U and M; U is the set of users and M is the set of messages posted during the observed time period. The colored nodes are seeds that help on community detection and how we use them will be explained later in this Section.

Semi-supervised strategy. In polarized debate, there are usually few users that are clearly biased towards one or more sides of a discussion, based on prior knowledge. For example, in a political discussion, the official profiles of candidates and parties are expected to only express opinions that are favorable to their side. We label users whose bias is clearly

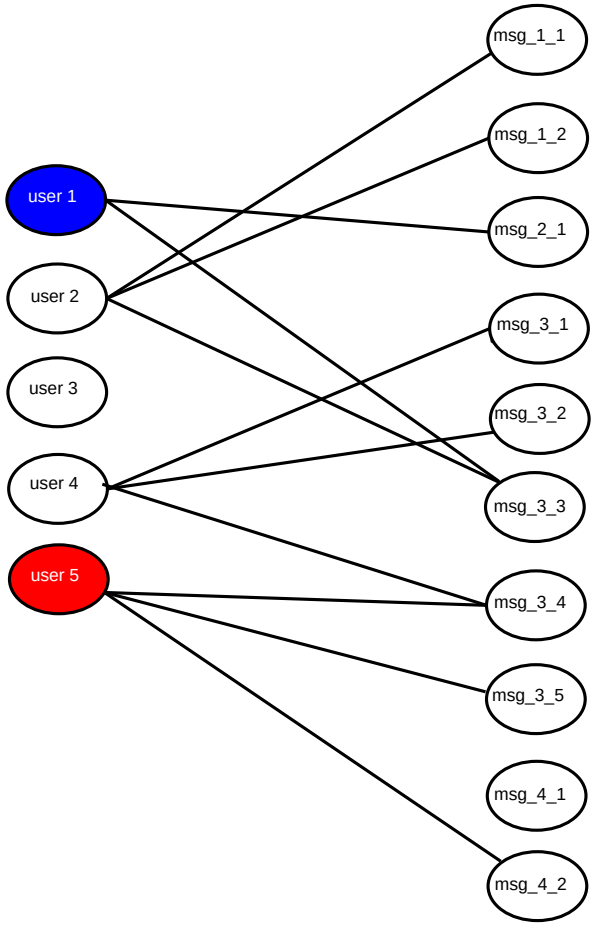


Figura 3.5. An hypothetical bipartite user-message graph. Nodes $user1$ and $user5$ are *seeds* that represent two communities that write messages with respect to topic T in a social network. Each message is identified with its author; for instance, message m_{3_4} is the fourth message posted by $user3$.

identifiable as representative of a particular side in a discussion as the seeds that will guide inference of class membership of the remaining nodes. We associate to each side k in topic T a set of positive seeds $S_k^+ = [s_{k1}, \ldots, s_{kj}]$, which are nodes that represent that side; formally, each user u is represented by $\vec{P_u}$, where $P_{ui} = 1$ if i is the side the seed represents, and $P_{ui} = 0$ otherwise.

In that direction, we can see the learning problem we are interested in as a *within-network classification problem*, a type of collective classification task where the network is *partially labeled*, i.e., ground truth labels are available for a subset of nodes [110; 51]. Within-network classification is important in several domains, such as image processing, classification of documents and fraud detection [40]. The goal of a within-network classification task is to use the network structure that connects nodes to infer the missing labels (recall that our labels are *soft*, since the polarity vector indicates a degree of membership

to each class/side). Within-network classification is a semi-supervised task: a learning algorithm will make use of both labeled and unlabeled nodes, propagating the knowledge obtained from labeled nodes to the unlabeled nodes through the network edges.

We will next detail how we use the graph G and the seet of seeds S^+ to infer the matrix P, by assuming that all edges are positive. Later, we will relax this assumption.

3.4.1 Propagating positive seeds through Random Walk with Restarts

For now we will leave aside the empirical observation from Section 3.3 that demonstrates how retweets can have a negative polarity and initially assume that all edges are positive. In that setting, our strategy to find communities around seeds will explore the rationale that entities of similar $P_{e,k}$ are likely to be close to each other in G. For example, in the toy example of Figure 3.5, $user3$ is closer to $user1$ than to $user5$ (through their common connection to message 3_3), what is an indicative that $user3$ is more likely to belong to $user1$'s community than to $user5$'s.

In fact, the notion of proximity of nodes in a network as evidence of node similarity has been applied in a wide range of problems, such as link prediction [100], collaborative filtering and content recommendation [65]. Many proximity measures have been proposed in the literature, ranging from the computation of the length of the shortest path between nodes to random walk based measures [153; 49]. We have chosen to adopt a proximity measure based on random walks with restarts, also known as *Personalized Page Rank*. Given a parameter α and a set of seeds S_k^+, Personalized Page Rank is defined for node similarity to seeds from side k is

$$PPR(\alpha, S_k^+) = \alpha * S_k^+ + (1-\alpha)pr(\alpha, s)W \tag{3.2}$$

where α is a constant in (0, 1] called teleportation constant, S_k^+ is a distribution called seed (or preference) vector, and W is the transition matrix. Equation 3.2 defines a Markov chain on nodes of G:

1. With probability α, the random surfer which is currently at node e follows a random edge which links to e.
2. With probability $1-\alpha$, the random surfer restarts at a seed uniformly chosen from set S_k^+.

We choose to employ random walks to find community around seeds for two reasons:

- Random walks have been successfully applied for the sake of community detection, for example, in MCL [156] and other approaches [6; 165]. The rationale is that, inside a community, a random walker will spend more time inside the community and will escape to other community with lower probability.

- A random walk is a stochastic process that outputs the probability that the random walker reaches a nodes, it thus fits nicely on probabilistic models and any algorithm that aims to compute outputs with degrees of confidence. Since our scenario carries an inherent uncertainty regarding the polarity of edges, working with probabilities makes it easier to reason about the nature of edges.

Given that it is previously assumed that the social graph is divided into K communities, we expect as input, for each community $k = [1, 2, 3, ..., K]$, a set of seeds $S_k^+ = [s_{k,1}, s_{k,2}, ..., s_{k,n}]$, that indicates which nodes (i.e., users or messages) are previously known to belong to community k.

For each set of seeds S_k^+, we run random walk with restarts, as in Equation 3.3. Each random walk propagates the known labels from the seed set to the remaining nodes – an strategy similar to *label propagation* semi-supervised approaches which have been successfully employed for classification in networked data [182].

$$rw_k^+ = RandomWalk(G, S_k^+, \alpha) \tag{3.3}$$

rw_k+ is a vector of dimensionality $||U|| + ||M||$; i.e., for each node it stores a probability > 0 that the random walker will pass through it. Based on rw_k^+ for $k = 1, 2, 3, ...K$, the probability that node e is positively leaned towards community k is defined as in Equation 3.4.

$$Pe, k = \frac{rw_{e,k}^+}{\sum_{i=1}^{K} rw_{e,i}} \tag{3.4}$$

For instance, assume K is 3 and $rw_{e,k=1}^+$ = 0.01, $rw_{e,k=2}^+$ = 0.001 and $rw_{e,k=3}^+$ = 0.05. According to Equation 3.4, $P_{e,k=1}$ = 0.164, $P_{e,k=2}$ = 0.016 and $P_{e,k=3}$ = 0.820, indicating that node e is highly positive leaned towards side $k = 3$, and negatively leaned towards sides $k = 1$ and $k = 2$.

In Figure 3.6 we show the communities, using colors, Pe, k for the toy example from Figure 3.5. In this example, two sides, blue and red, are represented by seeds $user1$ and $user2$. We then compute $rw_{e,blue}^+$ and $rw_{e,red}^+$ for each node e and $P_{e,blue}$ and $P_{e,red}$ according to Equation 3.4.

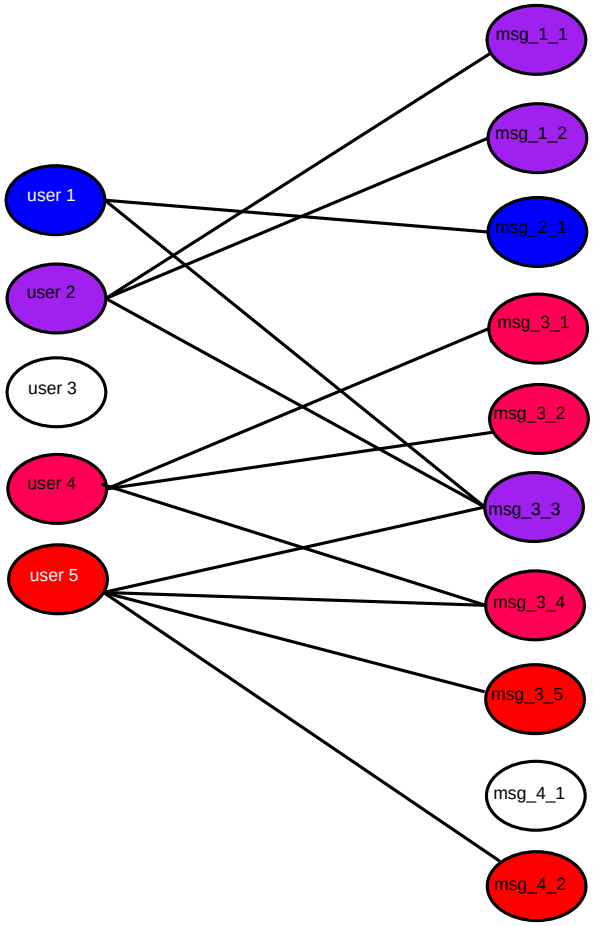


Figura 3.6. Computing rw^+ for the toy example from Figure 3.5. Intensity of red and blue denote the degree of membership to each node to each community.

3.4.2 Finding negative seeds with negative implicit feedback

As we demonstrated in Section 3.3, retweets and replies convey ambiguous signals regarding polarity, what makes inference of antagonism and support between communities a challenging task if one relies only on these interactions. Since edges can be negative, the random walker will likely traverse negative links, making antagonistic users and user-message pairs to have a high proximity in G, wrongly conveying the information that they share the same polarity.

In Table 3.3, we provide details on how retweets and replies can be conscious mechanisms of expressing either positive or negative interactions. In addition to retweets and replies, there is a third signal relating users and messages: absence of interaction. Intuitively, the fact that a user ignored a message with respect to him or her side can convey disagreement: since you do not agree with the content, is it reasonable that you do not share it to your network, or do not feel motivated to reply to it and argue against it. On the other hand,

absence of an interaction does not implicitly convey agreement. On average, we can then expect that the absence of an edge between a user and a message is more likely to convey antagonism than agreement.

interaction/polarity	**positive**	**negative**
retweet	endorsement	irony, sarcasm, quoting to criticize
reply	agreement	disagreement
no interaction	–	disagreement

Tabela 3.3. Retweets and replies can be either positive or negative interactions, while absence of interaction is not a mechanism of conveying positive sentiment with respect to a content. However, ignoring a content indicates an increased change of disagreement.

In fact, the exploitation of silence and absence of interaction as a useful signal to learn user preferences has already been explored with relative success in recommendation systems as a form of *negative implicit feedback*: most users do not provide explicit negative feedback on content they dislike; they simply do not consume it [101; 72; 177]. For example, most recommendation systems consider the buying action as an implicit positive signal, and the returning of a product as a negative signal [141]. [177] use dwell time – the amount of time the user spends visiting a page or item – as evidence to infer user opinion with respect to an item.

Based on prior work that successfully employed implicit negative feedback and on the intuition presented in Table 3.3, we devise a strategy to find a small set of messages that are negative with respect the entity they mention with a high probability, based on the less ambiguous signal of lack of interaction between users and messages, when compared to retweets and replies. More formally, based on our first approximation of user leanings represented on matrix P, we seek to build a set of seed messages S^-, where for $k = 1, 2, 3, ...K$, $S_k^+ = [s_{k,1}, s_{k,2}, ..., s_{k,n}]$. Each message m_i^{author} in S^- mentioning an entity which belongs to community k should satisfy three criteria:

1. message m_i^{author} should be ignored by users who are positively leaned to side k;
2. users who are positively leaned to side k should interact frequently with user $author$.
3. $author$ should be a popular user, to guarantee that lack of interaction is unlikely to happen due lack of visibility.

Condition 1 aims to capture the absence of interaction as evidence of disagreement, as summarized by Table 3.3. However, lack of interaction between a pair $(user, message)$ can still be ambiguous: $user$ can ignore $message$ either because he or she disagrees with it or because he or she is simply not aware of the message. Thus, we still suffer from the problem of ambiguity, but instead of positive-negative ambiguity, we now suffer from

ambiguity that, by relying just on Condition 1, we cannot be sure whether users are actively and consciously ignoring $message$ with a high probability. Our strategy to disambiguate between ignoring due to disagreement and ignoring due to not being aware of the message is to look for messages whose authors are frequently target of interactions by the community k; this means that many users follow and are aware of messages of $author$.

Equations 3.5 and 3.6 capture the intuitions from Conditions 1 and 2. Consider $E(M^{author})$ the set of edges arriving in messages authored by $author$.

$$P(\text{message from } author \text{ triggers reaction on community } k) = \frac{\sum_{u \in E(M^{author})} {P_{k,u}}^{+}}{||M^{author}||} \quad (3.5)$$

$$P(\text{message } m \text{ from } author \text{ triggers reaction on community } k) = \frac{\sum_{u \in E(M_m^{author})} P_{k,u}^{+}}{||E(M_m^{author}))}|| \quad (3.6)$$

An implicit feedback-based negativity score is computed as in Equation 3.7, by calculating the ratio of $P(usercommunity = k|author = a)$ and $P(usercommunity = k, message = m)$ and multiplying by two factors that account for $author$ popularity: $entropy(author)$ and the logarithm of $||M^{author}||$, the number of messages $author$ has authored. $entropy$ is calculated base on a vector that contains the counts of all $author$ interactions, by each interaction type. For instance, $[7, 10, 0, 3, 1, 2]$ represents that $author$ has been retweeted 7, 10, and 0 times by communities 1, 2 and 3, respectively. It also has had a message replied 3, 1 and 2 times by communities 1, 2 and 3. The intuition here is that high entropy values calculated from this interaction vector will denote users that are closer to community boundaries and interact with multiple communities. These usually will tend to be profiles of media outlets and influencers, whose opinions and posts are usually widely spread throughout the network.

$$implicit\_neg(m, k) = \frac{P(\text{community} = k|author = a)}{P(\text{community} = k|message = m)} * entropy(author) * log(||M^{author}||) \quad (3.7)$$

The set S_k^- can now be built by taking the top k messages of higher implicit negativity score for each community k. As in Equation 3.3, we then compute K random walks for each set S_k^-:

$$rw_k^- = RandomWalk(G, S_k^-, \alpha) \quad (3.8)$$

Finally, each element of matrix P is computed as in Equation 3.9.

$$Pe,k = \frac{rw^{+}_{e,k}}{rw^{+}_{e,k} + rw^{-}_{e,k}} \tag{3.9}$$

3.5 Case Study: Finding Communities and Community Relationships on Twitter

We use the strategy introduced in this Chapter to find communities and individual user and message polarities in online discussions about the Brazilian First Division Soccer League 2010/11/12/13/14 seasons. We chose official profiles of the soccer clubs as natural seeds that compose S^{+}. To run random walk with restarts, we set the teleportation ratio with the typical value of 0.85.

Figure 3.7 shows the distribution of implicit negativity score computed according to Equation 3.7. Axis x represent messages ordered by implicity negativity score, on logscale. Notice that a small set of messages (less than 100) exhibit a very high value of implicit negativity score, indicating that these messages are good candidates to serve as negative seeds. We chose to be conservative and use a small set of negatives seeds per community k of 10 seeds (i.e., messages). This choice is in accordance with recent studies that demonstrate that using few seeds can be an effective strategy for propagating labels in graphs [51; 102].

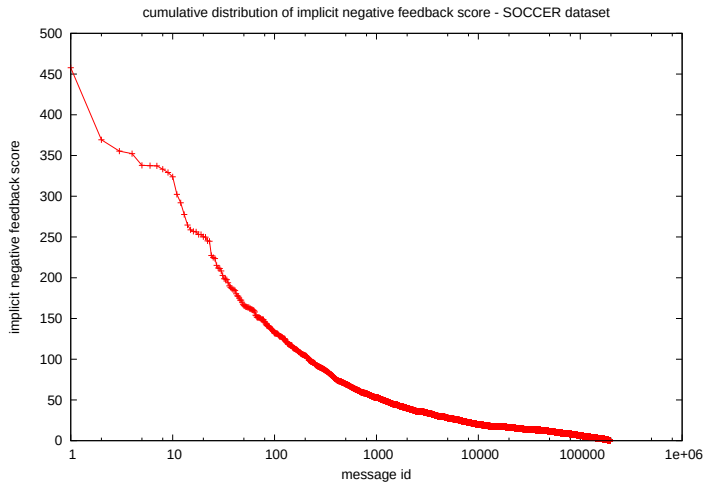


Figura 3.7. Distribution of implicit negative feedback score.

On Table 3.4 we show the average polarity of edges that cross each pair of communities. We show here only the top 12 values of lowest P. Notice that the pairs of communities with lowest average P tend to be the ones connecting supporters from the same Brazilian

Tabela 3.4. average P for edges crossing communities. In bold, the Brazilian state which is hometown to each team. Note that lower values of P (i.e., higher antagonism) is found on pair of communities representing supporters from the same state.

Group 1	**Group 2**	$average(P)$
Palmeiras **(SP)**	Corinthians **(SP)**	0.16
Atletico-MG **(MG)**	Cruzeiro **(MG)**	0.19
Santos **(SP)**	Corinthians **(SP)**	0.21
SÃ£o Paulo **(SP)**	Corinthians **(SP)**	0.22
Vasco da Gama **(RJ)**	Flamengo **(RJ)**	0.23
Internacional **(RS)**	Gremio **(RS)**	0.25
Fluminense **(RJ)**	Flamengo **(RJ)**	0.31
Vasco **(RJ)**	Corinthians **(SP)**	0.41
Vasco **(RJ)**	Fluminense **(RJ)**	0.43

state. We are able to correctly recover the local rivalries, correcting the inconsistency we showed in Section 3.3.

Figure 3.8 shows the cumulative distribution of Pm, k taken into consideration only messages which are targeted by retweets and only messages targeted by replies. As we expected, on average retweets tend to carry positive polarity more than replies: more than 80% of retweets come from users whose polarity with respect to the message is greater than 0.90.

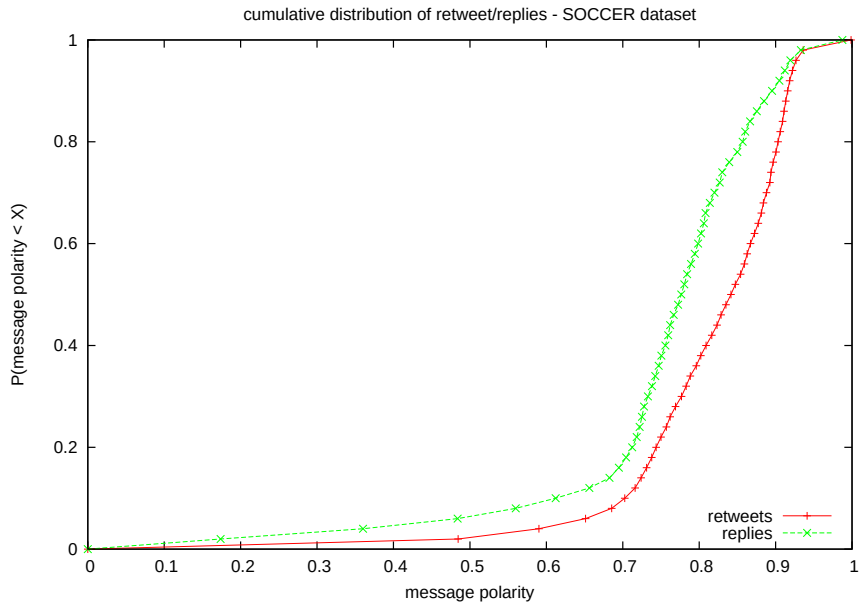


Figura 3.8. Retweets tend to be triggered by users who have positive view with respect to the message at a higher proportion than replies.

In Figure 3.9 we plot the cumulative distribution of how long a message has been retweeted after it has been originally posted by its author, measured in seconds. We plot this

distribution for 4 different cases:

- When a message w.r.t an entity is **positive**, but the user who retweeted has a **positive** polarity w.r.t this entity;
- When a message w.r.t an entity is **negative**, but the user who retweeted has a **negative** polarity w.r.t this entity;
- When a message w.r.t an entity is **positive**, but the user who retweeted has a **negative** polarity w.r.t this entity;
- When a message w.r.t an entity is **negative**, but the user who retweeted has a **positive** polarity w.r.t this entity.

We can observe from the Figure that retweets from user-message pairs which have a different polarity tend to occur more time after the original message has been posted, when compared to user-message pairs having the same polarity. We can see, for instance, that at least 30% of retweets in positive-negative and negative-positive cases occur after 16 hours of the original message posting time; on the other hand, on positive-positive and negative-negative pairs, only 10% of retweets occur so distant, in time, from the original post. Notice, also, that the four curves group into two clusters, corresponding to user-message pairs of same polarity and different polarity. This characterization demonstrates an interesting behavior on how social media users create new ways of using the social media system. While in isolation it is virtually impossible to tell whether a retweet is an endorsement or not, new signals captured from the social context, such as the “tweet reaction time”, can help on detecting irony and sarcasm, characteristics of human communication which are hard to detect by text itself [160].

3.6 Conclusions

Although a recent work has argued that the added value of negative links to the system is small [90], we do believe this might be the case only when positive links are unambiguous. In the case of networks were edges are not purely positive or negative, explicitly detecting negative relationships are important to correctly map community relationships.

Notice that the proposed model can be improved in several ways. The main opportunity of improvement is that, once we learn the negative relationships propagating random walks from the negative seeds, we can then refine our knowledge obtained from the initial set of random walks from positive seeds, avoiding traversing edges that have a high probability to be negative. We then can run a bayesian model that would do several iterations of this process until convergence.

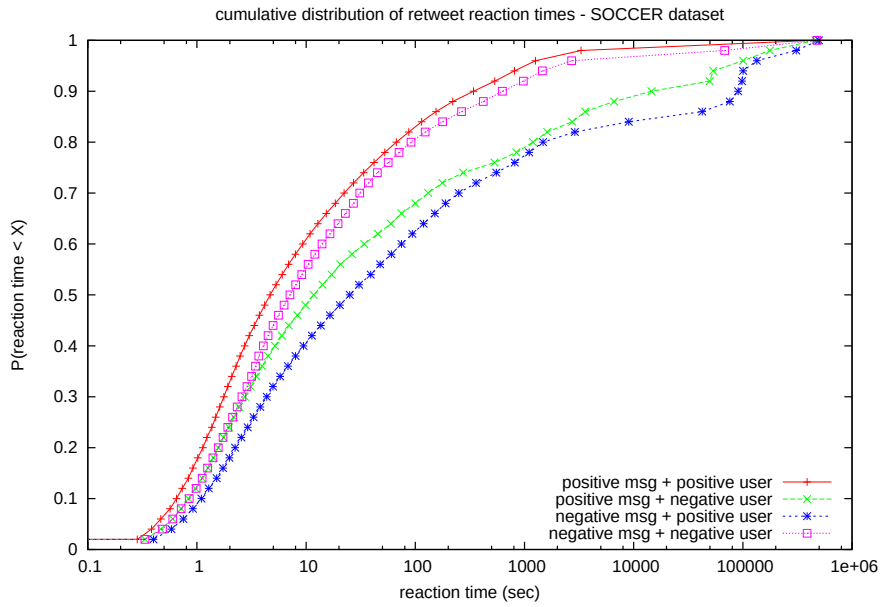


Figura 3.9. On average, retweets from users whose polarity is opposite to the message content tend to occur at later moments after the original message has been posted, indicating that the message is being broadcasted in a context that conveys irony or sarcasm.

In the next Chapter, we will focus on leveraging the division of social media users issuing opinions on a topic T into communities and associated community relationships to perform sentiment analysis in real-time, i.e., infer the sentiment of the textual content embedded in the M set, by exploring correlations between opinion holder bias encoded in P and sentiment expressed through messages in M.

Capítulo 4

Sentiment Analysis on Evolving Social Streams: How Self-Report Imbalances Can Help

As social media platforms become the primary medium used by people to express their opinions and feelings about a multitude of topics that pop up daily on news media [96; 27; 76; 20], the vast amount of opinionated data now available in the form of *social streams* gives us an unprecedented opportunity to build valuable applications that monitor public opinions and opinion shifts [75; 71], capturing instantaneous reactions of social media users and reflecting the buzz and dynamics of current happenings, breaking news and trends.

The ability to automatically distinguish positive and negative opinions on streams of opinion-based data supports many related web mining tasks in real-time, such as content recommendation and organization, search, user modeling and sentiment analysis. When a relevant event is taking place, offering mechanisms that allow users to navigate through opinions and monitoring the reactions of web users may enrich their web experience. For example, a political party may be interested in monitoring live reactions of the audience during a debate on TV, and feed the candidate with real-time feedback of what kind of opinions voters are showing [140]. Indeed, in the political scenario, it is increasing among social and computer scientists the belief that online social networks and social interactions influences political mobilization and actions [25], what suggests that the dissemination of positive and negative content about candidates and parties can actually influence voters' decisions. Another interesting application is to embed in a sports web portal a functionality that tracks the crowd sentiment during live matches, something far more appealing than the relative number of mentions to each team, which is what most sports web sites currently offer. Creating such applications enriches the personal experience of watching live events on TV, and following

the social media buzz simultaneously with live broadcasted events is becoming a multiple experience, where watching not only the event itself, but how others react to it, is part of the experience.

As discussed in Chapter 1, the task of interpreting positive and negative feelings expressed on social streams exhibits a number of unique characteristics that are not present in the static and well-controlled domains on which sentiment analysis has focused in the last decade – mainly product and movie reviews [155; 131; 71; 106]. On the downside, it faces two challenges that are common to many data stream classification tasks [111]: (i) the limited availability of labeled data and (ii) the need to deal with the evolving nature of the stream, which causes the target concept to change and requires learning models to be constantly updated – a problem known as *concept drift* [167]. Challenge (i) is a serious drawback because current sentiment analysis models are heavily based on supervised approaches [131; 155], and human constraints on generating a constant flow of labeled messages on streams remain high. The sparsity of language, the use of neologisms and word lengthening as an indicator of sentiment (e.g., "cooooooool!", "goooooooaal!" [28]) also contribute to make the process of acquiring large labeled sets of pre-classified messages unfeasible [71]. Challenge (ii) arises in sentiment streams as it is necessary to deal with constant changes of vocabulary and sudden changes of sentiment in reaction to real-world events. For example, in a few minutes a positive sentiment of the fans of a soccer team commenting on Twitter or Facebook may vanish by a goal scored by the adversary team; such *sentiment drift* represents a great challenge for real-time sentiment tracking, since it requires the stream classifier to be capable of quickly identifying and adapting to the sudden change on the dominant sentiment [143].

There are several challenges in performing such a task and the dominant approach relies on extracting textual patterns from messages and exploiting these patterns to predict polarity. Sentiment analysis and opinion mining[1] research have focused on the problem of classifying sentiment as a pure text classification problem. Different text classification algorithms have been applied to learn from word co-occurences and linguistic features to determine the sentiment contained in documents [76; 155]. Moreover, it has been used in conjunction with pre-defined lexicons to assess sentiment in political and movie review blogs [115].

Two broad categories of opinion analysis strategies can be identified in the literature: lexicon-based and classification-based [56] strategies. Lexicon-based approaches use lists of words containing positive and negative terms to compute the overall polarity of the document by counting the occurrence of those terms [155]. A clear disadvantage of this strategy is that lexicons are domain-dependent and the effort needed to generate lists of words may be high.

[1] Both terms are used interchangeably in the literature.

More recently, the increasing availability of opinion-based data in real time has motivated some studies that have analyzed sentiments in streaming data, specially over the Twitter microblogging system. Some approaches are as simple as the manual classification of tweets and lexicons of positive and negative words to monitor the debate performance of candidates in the 2008 U.S. Elections [41; 128]. While lexicons may provide sentiment analysis on an aggregated level, their coverage in terms of content is usually low because in complex contexts such as elections and sports, content is often ironic, contains subtle comments and refers to specific terms that only make sense at a specific time; it often lacks expressions of clear polarity such as “I love it” or “I hate it”.

Standard classification techniques have also been tested on Twitter [19], in addition to stream classification techniques such as the Multinomial Naive Bayes and the Stochastic Gradient Descent [20]. The major drawback of these approaches is that they require labeled data, which are very costly to obtain on a regular basis in a volume large enough to properly address concept drift. Active and semi-supervised strategies aim at reducing labeling and training efforts [30], but they still require training data to be sampled from a stationary distribution. In addition to that, in microblogs such as Twitter, the small document lengths restrict the possibility of using co-occurrence among terms and other standard text mining techniques to assign classes from an initial set of labeled documents.

In this work, we aim to explore the *opinion holder* as a critical aspect in understanding opinions on polarized debate. We aim to explore the biased nature of opinion holders (and consequently, opinions) on a polarized context to understand and process opinions on social media systems.

Indeed, the Linguistics field itself recognizes that context contributes to the meaning of textual sentences, expressions and opinions. There is a field of Linguistics – known as *Pragmatics* [118; 178] – which is dedicated to study the aspects of meaning and language which depends on the speaker, the target of the speech, the place and time where the conversation is taking place [99]. A classical example of an ambiguous sentence is “*Sherlock saw the man with binoculars*”, which can only be fully understood if more information about the situation is known.

On opinions on complex, polemic and heavily-debated issues, in some cases it is virtually impossible to interpret content without details of the broader context. We now discuss two examples of such dependence on opinion interpretation and context. On Figure 4.1(a), we show a screenshot of a YouTube video discussing a polemic event which took place during the 2010 Brazilian Presidential Elections – faced by three main candidates: Dilma Rousseff, Jose Serra and Marina Silva. Candidate Jose Serra, during the final weeks of the second round of the election, was hit by an object during a public protest – according to his partisans, he was hit by a hard and solid object; but, according to his oppositors, it was just a

smashed sheet of paper ("bolinha de papel", in Portuguese). Automatic inferring that a video mentioning this event and referring to "smashed sheet of paper" contains negative opinions and criticism on the candidate Serra is a very hard challenge to a text-based sentiment classifier, as "sheet of paper" is a term that, in most contexts, is not associated to any polarity in terms of positive or negative sentiment.

Figure 4.1(b) displays another YouTube video which highlights another major challenge in understanding opinions on broad and polemic topics: irony and sarcasm. This video was produced by partisans which are against the election of the current mayor of the Brazilian city of Belo Horizonte, Marcio Lacerda. However, the video's title claims that it contains "reasons for voting for Marcio Lacerda". Indeed, when watching the video, one quickly notices that such "reasons" are ironic and sarcastic. The context where the video is embedded – the set of people who posted and endorsed the video – is a key information to detect irony here, since it contains an unexpected opinion given what we know from the viewpoints of the users who generated the content [160].

(a) Reasons for "voting" on a candidate (b) The "Bolinha de papel" Case

Figura 4.1. Ambiguous and complex opinions on Politics expressed on YouTube videos.

Previous work on Sentiment Analysis has already highlighted the context-dependency of the relationship among words and topics; [32] points out that "NASDAQ up is accelerated" and "Global temperature up is accelerated" are two sentences with different polarities (positive and negative , respectively) and whose interpretation depends on the topic and context.

In despite of ambiguity, debate on complex topics such as Politics, Sports and Public Issues tend to rely on more subtle and complex aspects of language expression; thus, sets of affective words (words that express feelings, such as "satisfied" [77]) and evaluative words (such as "good" and "bad") are not likely to cover a satisfatory proportion of the opinions.

Despite these important constraints and drawbacks, streams reflecting the society's immediate emotional reactions regarding a topic have an important property, which we seek to

exploit in this work, namely, the flow of opinions from social networking services is inherently constrained to manifestations from individuals that have explicitly and deliberately *chosen* to post a message in reaction to some real-world event; thus, the distribution of positive and negative opinions is potentially quite different from the random samples obtained in traditional opinion polls and survey methodologies [104]. Although such *reporting bias* is usually perceived as a source of inaccuracy [82; 54], here we argue that the self-reporting nature of social media, when observed on large-scale social network data, may actually provide signals that ease the task of sentiment tracking in online environments, provided that we understand the **factors** that motivate people to publicly express their feelings. We build sentiment analysis models that exploit two factors widely described by substantive research from social psychology and behavioral economics that describe human preferences when disclosing emotion publicly:

> **Positive-negative sentiment report imbalance:** People tend to express positive feelings more than negative feelings in social environments [120; 18; 42; 94; 78].
>
> **Extreme-average sentiment report imbalance:** People tend to express extreme feelings more than average feelings in social environments [7; 39; 38; 82].

We explore each of these two self-report imbalances to accomplish a different subtask in learning-based sentiment analysis. The first self-report factor, which we call **positive-negative sentiment report imbalance** throughout this chapter, is employed to acquire labeled data that supports supervised classifiers. In the context of *polarizing groups* – a division of the population into groups of people sharing similar opinions in the context of a topic [12; 64], a positive event for one group tends to be negative to the other, and vice-versa. For example, while supporters of a football team are likely to be happy when their team scores, fans of the adversary team are expected to be upset when faced with the same event. Based on social psychology research that states that the disclosure of positive feelings is preferred, we can then make a prediction of the current dominant sentiment by simply counting how many members of each group, relative to group sizes, decided to post a message during the specified time frame. Since the social context information only holds during time frames when a significant real-world event happens, we adopt a probabilistic model that computes the uncertainty of the social context, and, at each time frame, generates a probabilistic sentiment label, which can then be incorporated into a range of content-based supervised classifiers.

The second self-report factor we explore is related to the human tendency to report extreme experiences more than average experiences [7; 39; 38; 82]. The **extreme-average sentiment report imbalance** implies an important consequence for real-time sentiment tracking: because extreme feelings stimulate reactions, spikes of activity in streams of opinio-

nated text tend to contain highly emotional terms, which are precisely the features that are helpful for sentiment prediction. We propose a simple text representation strategy based on this observation, named *term arousal*, that maintains, for each term (or lexical unity, e.g., n-grams), a measure of how often it appears in high-volume time windows in the stream; we call these **high-arousal** terms. Our experimental studies demonstrate that these terms are better indicators of emerging and strong feelings than traditional static representations (e.g., TF-IDF), allowing the underlying classification model to adapt quicker to sudden sentiment drift induced by real-world events.

In summary, our main contributions in this Chapter are:

1. We raise awareness over the fact that opinions expressed on social media platforms are *not* a random sample of the online population, but are impacted by many social and psychological factors that need to be accounted for in order to build reliable and useful sentiment analysis systems;
2. We show that, in the context of online polarized discussions, self-report imbalances create rich *social contexts* that can be leveraged to improve two key subtasks in the construction of a sentiment stream classifier – namely, the acquisition of labeled data and feature representation suitable to deal with sudden sentiment drifts.

We evaluated our social psychology-inspired framework on sports events heavily debated on Twitter; when instantiating our framework with a SVM and Multinomial Naive Bayes classifier, our results are comparable to what is typically obtained as an acceptable result for document-level sentiment analysis – between 80% and 85% of accuracy [155] – but, because the stream-based scenario imposes stricter and harder constraints, we believe they point to a promising option for sentiment classification on evolving social streams. In addition, our approach targets two generic sub-tasks for learning-based sentiment analysis – label acquisition and feature representation. As a result, our framework can be incorporated into sophisticated sentiment classifiers that make use of more powerful NLP models and features.

4.1 Social Psychology Background

Psychologists classify emotions into two independent dimensions: pleasure (happiness or sadness) and activation (or arousal) [14; 15], as shown in Figure 4.2.

We aim to explore in this Chapter the phenomena, widely observed on social psychology literature, that emotions disclosed on social environments are biased toward the positive and high arousal extremes of the bidimensional space show in Figure 4.2, having a disproportional ratio of feelings such as *excited* and *elated*. The *self-report imbalances* we briefly

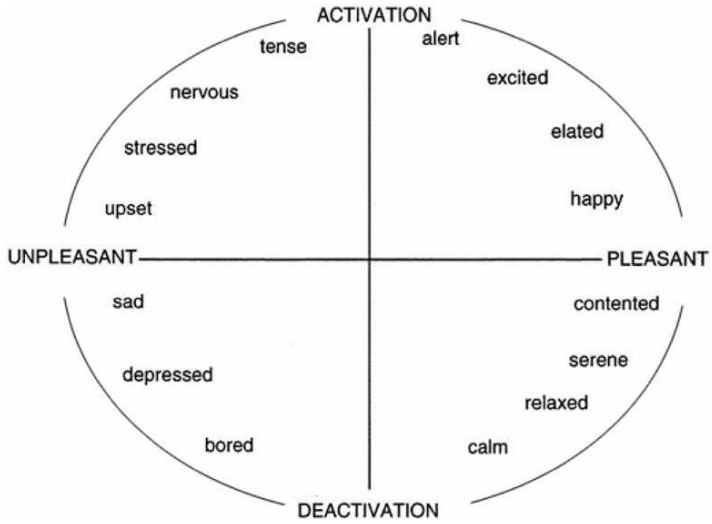


Figura 4.2. Semantic structure of affect and emotions [14]. Emotions are classified into two dimensions: pleasure and activation (arousal). For example, calmness is a a low-activation/arousal and neutral (not pleasant nor unpleasant) emotion; a nervous person is experiencing a high-activation and unpleasant emotion.

presented earlier in this Chapter are biases in this bidimensional emotion space caused by the fact that social media systems are *communicative* platforms; as a consequence, opinions and feelings expressed in online social environments are a result of opinion holders' explicit desire to make his friends or followers aware of his or her opinions. In other words, the communicative nature of social media makes social data a side effect of intentional and deliberate communication between users, rather than a representation of some underlying activity [137; 104].

On the positive-negative dimension, the preference on the disclosure of positive feelings is caused by our need for being perceived as successful and happy persons [116; 138], and it causes a bias where everyone in online social environments perceives others as happier than they actually are [78]. It has been recently found that private messages in social media tend to contain proportionally more negative messages than public messages [16].

In the case of opinions expressed over a polarizing topic, the preference on sharing positive news and opinions goes beyond the human's desire to improve his or her reputation: each group also gives preference to news and facts that favor their viewpoints, a result of many biases such as *confirmation bias* and *selective exposure* [133; 104]. A recent report from Pew Research Center, for instance, showed that 52% of Americans declared themselves happy with President Obama's reelection in 2012, but a sentiment analysis on Twitter unveiled that 77% of Twitter users felt the same way [120].

Notice that the definition of a *positive* event is group-dependent: for rival supporters of a team or opposers of politicians in office, negative facts such as a conceded goal or a

political scandal will be explored by them as "positive" – i.e., as a motivation to explore the fact to their benefit. Also, in some contexts, such as product reviews, the bias leans toward the disclosure of negative experiences [70]; our sentiment analysis framework is generalized to take advantage of the asymmetry on either direction.

On the activation (arousal) dimension, it was found that extreme emotions – angriness, anxiety, awe, excitement – are *high-arousal* emotions: they affect our body and put us in a state of activation and readiness for action [15; 18]. In social media, action means making private feelings public, what makes sentiment expressed on online media to be biased towards strong feelings and opinions.

In the next sections we will detail how we embed these biases on sentiment self-report in the analysis of feelings expressed on social streams on polarized debate.

4.2 Positive-Negative Self-Report Imbalance

Differently from the majority of research on supervised sentiment analysis, which focus on batch processing of opinionated documents [131; 155], here we are interested in the setting where the data arrives as an infinite stream and reflects real-world unpredictable events. As we previously, in this setting a constant flow of labeled messages is required to build and update supervised sentiment models. Unfortunately, in textual streams characterized by sparse and time-changing content it is not feasible to manually obtain labeled data in significant amounts and in a timely manner [111].

To overcome this problem, we propose a method to acquire labeled messages by exploiting the *positive-negative sentiment report imbalance* in the context of polarizing groups. On Chapter 3, we detailed our solution that estimates the preference B_u of each user toward a set of monitored entities in a polarized debate. We use knowledge from polarization, bias and self-report theoretical and empirical studies to assume that biased users will privilege posting positive messages that favor their own opinions and viewpoints.

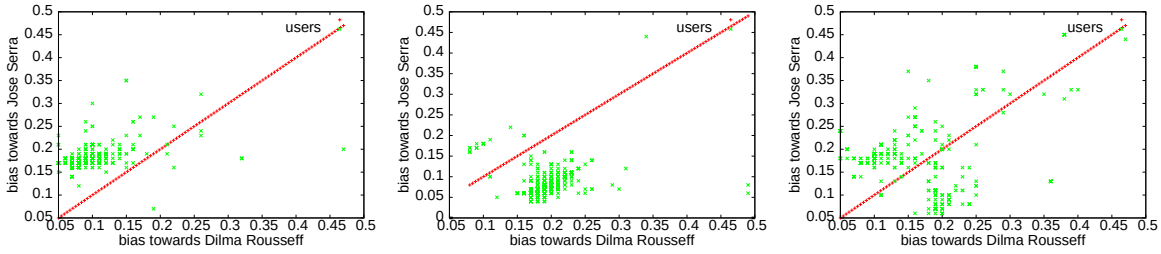
Propagating bias across terms. We transfer information from users to terms by assuming that term t will be *positive* toward entity e if it is adopted more frequently by users biased toward entity e than by users of different sides in tweets that mention e. Similarly, t will be *negative* to entity e if it is adopted by a large number of users who oppose that entity, in contrast with the number of supporters of e. Neutral content is expected to be endorsed by both sides. To validate this intuition, in Figure 4.3 we plot the bias vector associated with users that referred to three different web pages in their tweets: a YouTube video with positive comments about Jose Serra (Figure 4.3(a)), an official video from Dilma Rousseff's campaign (Figure 4.3(b)), and a general news article about the Brazilian 2010 Presidential

Elections (Figure 4.3(c)). We can understand each dot in Figure 4.3 as a vote for a label (positive/negative) of the content. Note that this computation *propagates* user bias information to all messages that contain at least one term adopted by a user with known bias. As user bias does not change often and tends to be consistent over a period time for most users, we can deal with the nonstationary nature of social stream: new terms may arise and old terms may change their meaning (**Challenge 2**), but users keep providing reliable judgments.

In order to transform user bias into term polarities, we take into account the bias vector associated with each user that used term t. A possible unsupervised approach is to compute the sum vector of the polarity vector of all users that refer to entity e by adopting term t:

$$\vec{B}_{t,e} = \sum_{u \in V} \vec{B}_u \tag{4.1}$$

Note that this computation *propagates* user bias information to all messages that contain at least one term adopted by a user with known bias, thereby revealing the judgement of the content produced by users with unknown bias. This is important because it is expected that information on user bias will be available for only a portion of users, since many users are never involved in endorsement interactions.



(a) positive term for Jose Serra (b) positive term for Dilma Rousseff (c) neutral term

Figura 4.3. User bias vectors for three different contents, which show that user bias is a good predictor of term polarity – 2010 Brazilian Presidential Elections.

Dealing with concept drift. When a term is adopted for the first time, it will have the same bias as the corresponding user who adopted it. As new messages pass through the stream, $\vec{B}_{t,e}$ is updated incrementally. As such, users collectively judge new terms, referring to them (or not) in their messages. To predict the polarity of a message d, we first convert the bias vector of each term present in a message into *polarity probabilities*. Given the bias vector $\vec{B}_{t,e}$, and that $\vec{B}_{t,e,e}$ represents the strength of component e in $\vec{B}_{t,e}$, we calculate the probability that term t refers positively to entity e according to Equation 4.2.

$$\hat{p}(polarity = +|t, e) = \frac{B_{t,e,e}}{||\vec{B}_t||} \tag{4.2}$$

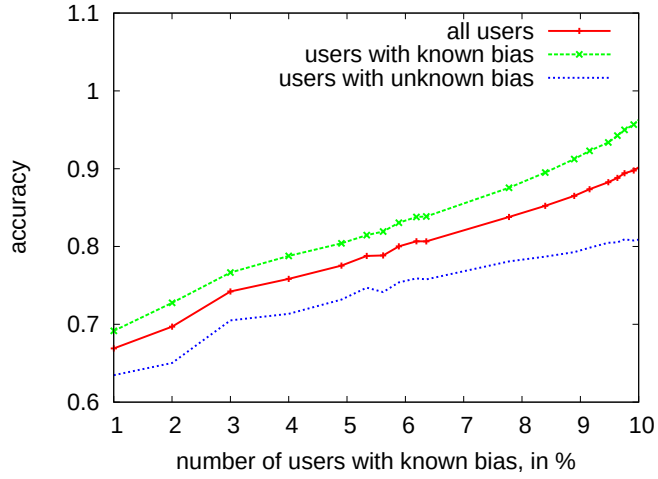


Figura 4.4. F1-accuracy level for different ratios of users with known bias – 2010 Brazilian Presidential Elections Twitter Dataset. As the ratio of users of known bias increases, the F1-measure increases, even for tweets posted by users with unknown bias.

Note that we compare the strength of bias to e in $\vec{B}_{t,e}$ with the magnitude of the bias vector. Specifically, $\hat{p}(polarity = -|t, e)$ may be calculated as 1 - $\hat{p}(polarity = +|t)$. To predict message polarity, we may adopt various strategies to combine those probabilities. Limited to 140 characters, Twitter messages are short, thus, we exploit a simple strategy for predicting message polarity, which is to consider the term of highest polarity in each tweet: $polarity = argmax(\hat{p}(polarity = x|t))$.

In Figure 4.4, we analyze the performance of our transfer learning approach as the fraction of users whose known bias varies. We report performance numbers using the $F1$ measure. To generate ground truth with respect to messages, we combined manual labeling with automatic labeling for messages containing tags that clearly indicated a preference for a specific entity. To make our evaluations fair, we removed all tags used to generate our labels from message content. We can see that the F1-measure increases as the ratio of users with known bias increases, up to a value at which F1 stabilizes. When the bias of 15% of users commenting on politics is known, the F1-measure equals 85%, while in the corresponding case for soccer, F1 is 90%. Note that the F1-measure for posts from users with unknown bias also increases as we transfer bias from a greater number of users, what further demonstrates the applicability of our user–term bias transfer approach.

Comparison with SVM. We now compare the F1-measure provided by our bias-based sentiment analysis model against the same metric provided by a typical SVM classifier. We chose SVM because it has already been successfully applied to various sentiment analysis application scenarios, including the analysis of tweets [19; 155]. In order to perform this experimental comparison, we split each dataset into two partitions. The first partition is used

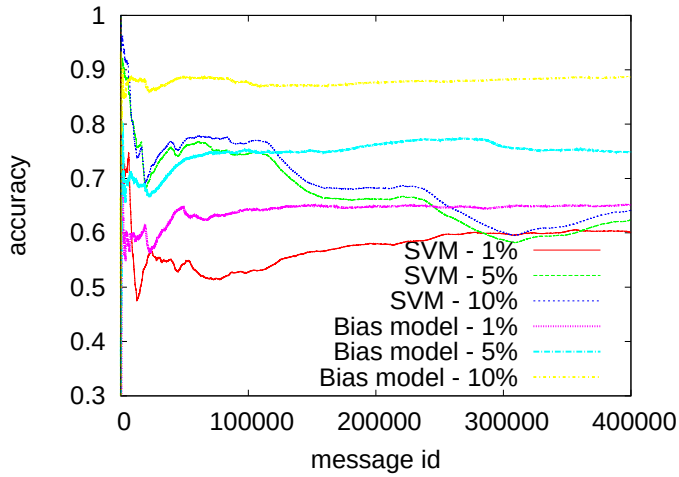


Figura 4.5. Bias-based model versus SVM classifier – 2010 Brazilian Presidential Elections Twitter Dataset

for training, and it comprises the first 10% of tweets from each dataset. The second partition is used to validate each approach. Our comparison involves the execution of different training configurations. More specifically, each execution uses 10%, 50% or 100% of the training partition. For SVM, the training partition was used to train textual-based models, while for our bias-based model, we only considered endorsements in order to compute the OAG and generate bias assessments for users. When we compared the results on a chronologically ordered set of labeled tweets (i.e., the test partition), as shown in Figure 4.5, some important observations arise. We can note that the SVM F1-measure decreases across time, which is evidence of changes in the textual feature distribution. In contrast, the bias-based sentiment classifier is able to mantain a stable F1-measure, as it incrementally incorporates bias information on new terms by propagating user bias.

4.2.1 Temporal Positive-Negative Self-Report Imbalance

We can make better use of positive-negative report imbalance by observing differences on the strength of reactions of polarization groups during a specific time span, moving from processing individual messages to processing groups of messages. These groups are obtained by dividing the social stream into a sequence of non-overlapping and contiguous time windows of equal duration (e.g., Δt minutes), what gives us the capability of exploiting the **social context** induced by the set of users that expressed their sentiment w.r.t. topic T during each time window W_t. Each window W_t contains all messages sent during the time period $[t_i, t_i + \Delta t]$ (W_0 starts at t_0 and $t_{i+1} = t_i + \Delta t$) and is composed of a triple (S_t, D_t, Y_t):

- S_t is a multiset of group memberships of all users who posted a message during W_t.

On a polarized domain, we assume that each user belongs to one of two groups, G_A or G_B.[2] For instance, $S_t = \{G_A, G_A, G_A, G_B, G_B\}$ indicates that 3 members of group G_A and 2 members of group G_B posted a message during W_t. Assigning users to groups is a task that can be accomplished by several community detection and graph mining techniques that explore the social ties among users, under the assumption that similar users are likely to connect to each other [3], here we used the Personalized Page Rank strategy presented on Chapter 3 and we take the largest component of the bias vector as the user group.

- D_t is the sum vector of all feature vectors extracted from messages written during W_t;
- $Y_{e,t} \in \{+, -\}$ indicates the ground-truth sentiment expressed during W_t w.r.t. an entity e in the context of topic T. Here, each e is an individual or organization naturally linked to the polarizing group that supports it; for instance, if $G = \{\text{Democrats}\}$, than $e(G) = \{\text{Barack Obama}\}$, and $e(G) = \{\text{New York Giants team}\}$ if $G = \{\text{New York Giants fans}\}$.

Note that, instead of seeking for labels for individual messages, we label *all* the messages mentioning an entity e in time window W_t with the same polarity $Y_{e,t}$. Although we do not expect every opinion expressed during a time window to follow the same polarity, we seek here to determine the *dominant* sentiment during W_t; furthermore, the probabilistic method we will detail next assigns a confidence on the label estimation, what can be interpreted as an estimate of the proportion of positive and negative messages written during a given W_t.

For now we ignore the content vector D_t and focus on S_t as an input to build a sentiment prediction function $f : S \rightarrow Y$. The fundamental principle we seek to exploit is that, on polarized discussions dominated by two opposing groups G_A and G_B, in general $Y_{e(G_A),t} = +$ implies that $Y_{e(G_B),t} = -$, and vice-versa (we will relax this requirement in Section 4.3, by learning a content-based classifier based on labels provided by S_t). A simple approach to predict Y_t based on S_t is to consider that each message is a "vote" toward the sentiment expected to drive more reactions and, thus, a majority-voting strategy is employed to predict the dominant sentiment at W_t. In the toy example $S_t = \{G_A, G_A, G_A, G_B, G_B\}$, since we are supported by social theories that indicate preference toward the report of positive sentiment, we would predict 3 votes for labels ($Y_{e(G_A),t} = +$, $Y_{e(G_B),t} = -$) and 2 votes for labels ($Y_{e(G_A),t} = -$, $Y_{e(G_B),t} = +$). The only point of caution here is that normalizing by group sizes $|G_A|$ and $|G_B|$ is important to discount the effect of larger groups on S_t.

[2] In practice, a domain can be associated with more than two groups, i.e., N=20 groups of supporters are found on National Football League. However, at each event of interest (e.g., a football match), we focus on the two polarizing groups that have a direct interest on it.

Majority-voting is a simple and straightforward approach, but it has an important limitation: it does not quantify the uncertainty on the information provided by the voters [142]. Since the labeling mechanism by social context is not perfect, capturing the degree of confidence on the correlation between S_t and Y_t is crucial if we will incorporate this information on learning models. In particular, the labeling scheme based on positive-negative report imbalance is error-prone due to two reasons:

1. S_t is likely to carry a significant correlation with the dominant sentiment only when a well-determined and relevant event happened during time window W_t, i.e., a goal or touchdown in a sports match, or some breaking news on the topic being followed. Most of the time, the positive-negative report imbalance will not be triggered at a sufficient strength, and an unreliable prediction will be generated.

2. Since we are modeling only user posting decisions in face of positive/negative events and abstracting from several other factors that influence the posting decision (as well as different individual posting probabilities), we are prone to deal with noise due to the many factors that motivate user reactions and that we are not accounting for.

Therefore, in order to make our approach reliable and more useful, it is desirable to associate with each predicted label Y_t a measure of confidence $P(Y_t|S_t)$ that captures the noisy nature of the multiset of group memberships S_t. We instantiate a probabilistic model that assumes that on each time window W_t a coin of bias θ_t is tossed to decide whether each message will be authored by a member of G_A or G_B, and $|G_{A,t}|$ messages from members of G_A and $|G_{B,t}|$ from members of G_B are observed. A fair coin is expected to generate a number of heads (G_A) and tails (G_B) proportional to $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$ and $1 - \theta_{fair}$, respectively, modeling the fact that members of both groups are reporting their sentiment with the same probability. Alternatively, a biased coin, whose θ_t is different from $\frac{|G_A|}{|G_A|+|G_B|}$ at some degree, means that members of one group are self-reporting their feelings at a higher rate than the other, indicating that its members are probably experiencing positive feelings in comparison to the other group.

A coin model is convenient because it naturally models the intuitive fact that spikes of activity in the social stream are more informative: in the same way that our confidence on the bias of a coin increases as we toss it more times, a time window W_t which contains a large number of messages (and, consequently, a larger multiset S_t) is more likely to carry a clear dominant sentiment, not only due to a larger sample, but because spikes of activity are likely to be associated with real-world events that trigger the positive-negative report imbalance. Our probabilistic model is divided into two steps:

1. Estimate the probability distribution on the latent variable θ_t;

2. Estimate how far θ_t is from $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$.

We use Bayesian estimates in both steps. To estimate the uncertainty on θ_t, we need to calculate the posterior predictive distribution $P(\theta_t|S_t)$, i.e., the distribution over θ_t after observing the resulting multiset S_t. In Bayesian inference, the posterior $P(\theta_t|S_t)$ is proportional to a likelihood function $P(S_t|\theta_t)$ and a prior distribution $P(\theta_t)$; we adopt the classical Beta-Binomial model: $P(S_t|\theta_t)$ is computed from a binomial distribution $Bin(|W_t|, \frac{|G_{A,t}|}{|G_{A,t}|+|G_{B,t}|})$ and the prior follows a Beta distribution $Beta(a, b)$ (a and b are hyperparameters) [142; 22]. As a result of the conjugacy property of the Binomial and the Beta distributions, the posterior predictive distribution nicely follows a Beta distribution $Beta(|G_{A,t}|+a, |G_{B,t}|+b)$ that captures our uncertainty over θ_t [22].

It is still necessary to choose the hyperparameters a and b that govern the prior distribution $P(\theta_t)$ and capture the knowledge acquired from previous observed data streams over the noisy nature of the coin. To incorporate our prior knowledge that θ_t is expected to be proportional to group sizes, we want to find hyperparameters a and b in the form $a = \frac{K|G_A|}{|G_A|+|G_B|}$ and $b = \frac{K|G_B|}{|G_A|+|G_B|}$. K can be understood as a smoothing parameter: the greater its value, the more confident the model is that θ_t is close to θ_{fair} and less importance will be given to the data. On the other hand, if we choose an uniform prior $Beta(1, 1)$, then we let the model rely totally on the observed data to judge how likely the tosses are coming from a coin of bias θ_t; the expected value of the coin bias in this case is equivalent to the maximum likelihood estimate $\theta_t = \frac{|G_{A,t}|}{|G_{A,t}|+|G_{B,t}|}$ [22]. Such direct estimation of θ_t makes the unrealistic assumption that tosses are generated i.i.d. from a noiseless coin.

We estimate K from the streaming data by employing an Empirical Bayes approach[3]. To learn the extent to which the coin we are modeling is noisy, we take advantage of the data continuity in the stream: we observe a sequence of noisy estimates $(\theta_0, \theta_1, ..., \theta_t)$ of a different coin being tossed at each time window. The property we want to explore here is that we expect consecutive time windows W_i and W_{i+1} of *similar message volume* to share a similar θ; large differences in θ between these windows should be attributed to noise, since no significant real-world event has happened (otherwise we would observe a large $||S_{i+1}| - |S_i||$). On the other hand, we would like to allow consecutive time windows with a large difference in message volume to exhibit a larger absolute difference $|\theta_{i+1} - \theta_i|$, since, according to our user behavior model, a spike of activity will trigger a bias either on G_A or G_B.

We seek to find the value of K that maximizes Equation 4.3. ρ is the Pearson correlation coefficient, and ΔV and $\Delta\theta(K)$ are vectors containing the sequence of $||S_{i+1}| - |S_i||$

[3]Empirical Bayes methods are approaches that estimate the prior distribution over a random variable from the data itself, rather than defining the distribution before observing any data, as on standard Bayesian inference [55].

and $|\theta_{i+1} - \theta_i|$ observed on the stream. Note that we write $\Delta\theta(K)$ as a function of K, since the estimates of θ_t are affected by the prior distribution $P(\theta_t|K)$. The highest Pearson correlation will explain larger differences in θ through larger differences in time-window volume, and we estimate it by using a standard gradient descent method.

$$K = argmax(\rho(\Delta V, \Delta\theta(K))) \tag{4.3}$$

Recall that our goal is to estimate how far the latent variable θ_t is from $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$, what indicates a bias in the posting decision of either G_A or G_B. This value can be estimated by calculating the area under the curve of the distribution $Beta(|G_{A,t}| + a, |G_{B,t}| + b)$ at the decision threshold $x = \frac{|G_A|}{|G_A|+|G_B|}$. If $I_x(a, b)$ is the CDF of $Beta(a, b)$ in the interval $(0, x)$, then

$$\begin{aligned} conf(\theta_{fair}, S_t) = \max(&I_{|G_A|/(|G_A|+|G_B|)}(|G_{A,t}| + a, |G_{B,t}| + b), \\ &1 - I_{|G_A|/(|G_A|+|G_B|)}(|G_{A,t}| + a, |G_{B,t}| + b)) \end{aligned} \tag{4.4}$$

where I is the regularized incomplete Beta function and can be used to determine to the cumulative distribution function in a Beta distribution [142]. The value $1 - conf(\theta_{fair}, S_t)$ gives us an estimate of how likely the predicted label is trustable given the observed social context S_t, i.e., $P(Y_t|S_t)$.

4.2.2 Experimental Evaluation using Twitter data

We evaluate the predictive power of social contexts induced by the positive-negative report imbalance on the analysis of the reactions expressed on Twitter by fans of two popular sports that generate passionate debate on social media: soccer and (American) football. Sports competitions are among the topics that generate the largest fractions of audience both in broadcasting media [166] and social media [93]; however, most initiatives taken by content portals to turn the live game experience into an online social experience are still restricted to simple tools such as the display of the most popular tweets or plots on the variation of the relative number of mentions of the playing teams. Measuring the crowd sentiment during live matches is something far more appealing and may answer relevant questions such as "do the supporters still believe in a win, despite losing the match so far?".

Table 4.1 gives an overview of two datasets we obtained from the Twitter data collection API. The datasets comprise fans' debate on Brazilian Soccer League seasons (2010, 2011 and 2012) and NFL (2010/11, 2011/12 and 2012/13 seasons). We chose team names and specific words of each competition as keywords. More than 35.8 million tweets from

5.6 million users have been collected in the SOCCER dataset, and 23 million tweets from 4.2 million users in the case of the NFL dataset. While tweets on Brazilian soccer are mostly in Portuguese, NFL debate is dominated by English, what gives us the possibility to experiment our model in two languages, after we build a content-based stream classifier in Section 4.3.

Tabela 4.1. General overview of the datasets collected from Twitter.

	Soccer	**NFL**
seasons	10-11-12	10/11, 11/12, 12/13
language	Portuguese	English
# of user groups (teams)	12	20
# of tweets	35,834,453	23,094,280
# of users	5,638,906	4,230,731
# of users w/ 1+ post/week	35,121	58,981

Before performing any sentiment prediction, we need to segment the user base into polarizing groups. In the sports domain, the natural criterion for dividing users into polarizing groups is to reflect their team preference. Several community detection and graph mining approaches that leverage social ties and social interactions can be used to accomplish this task; we manually labeled a set of users with their team preference and then used the similarities in their retweet pattern to estimate the class of unlabeled users, according to what we detailed in Chapter 3.

Due to the highly-dynamic nature of sporting events, we analyze sentiment and social contexts in 1-minute time windows; larger time frames may be suitable for less dynamic domains. To generate ground-truth sentiment labels, we examined the match facts and the evolving sentiments for a number of matches in the SOCCER and NFL dataset. In addition to the match score, we manually examined the content of tweets and also included cases where the match score did not reflect the sentiment, as soccer matches that ended as null ties (0–0), but the result was enough to grant one of the teams the championship title. Although each time window is associated with a set of messages, we aim to determine the overall, global sentiment which dominates each time window, instead of individually trying to predict the polarity associated with each post.

Figure 4.6 shows the accuracy on the sentiment prediction task for the two datasets. On the x axis, we grouped time windows according to its volume in relation to the average time window volume: $bin = i$ corresponds to time windows where the number of messages were between i and $i + 1$ times the average.

We observe that, for high-volume time windows, accuracy is very high: we could predict with more than 90% of accuracy the dominant sentiment on time windows whose volume of tweets were at least 5 times the average, despite not taking any textual content into account. This result validates the sociopsychological principle that motivated our method –

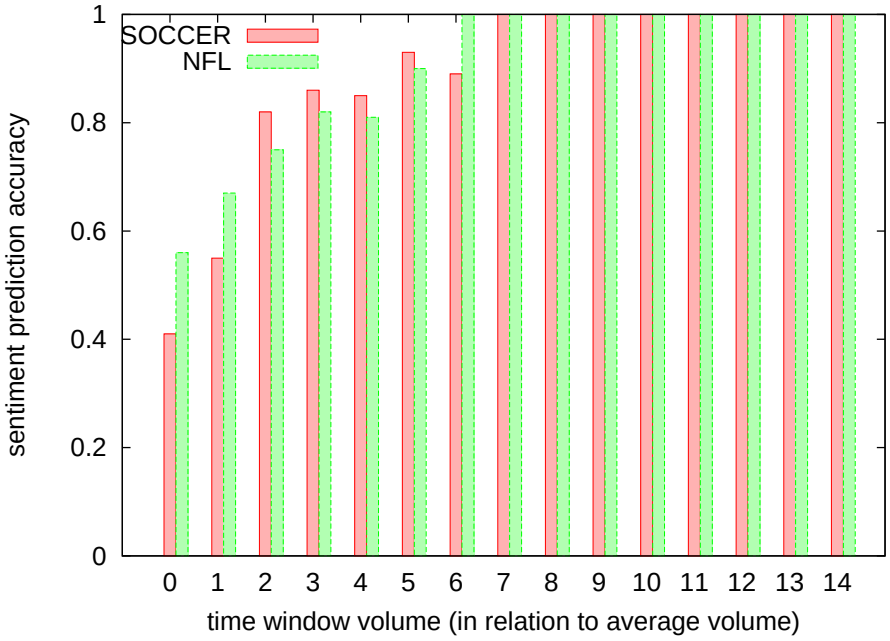


Figura 4.6. Accuracy on sentiment prediction on 1-minute time windows. Ground-truth was established by manual examination of a sample of tweets in each interval; we grouped time windows according to its volume in relation to the average time window volume. Social contexts based on positive-negative sentiment report imbalance are highly effective on sentiment prediction on large-volume time windows: for time windows whose message volume is higher than 7 times the average window volume, accuracy is practically 100%.

positive and negative feelings are disclosed with different probabilities – and, confirms that, in the sports domain, sentiment report is biased toward the positive feeling.

We can also note from the histogram that accuracy decreases with the volume of tweets in the time-window; on time-windows whose volume is above average, accuracy is comparable to a random guesser, meaning that the induced social context is not relevant and the positive-negative report imbalance is not triggered in sufficient strength, and other factors are affecting the posting decisions' of members of G_A and G_B.

Since the majority of the time windows are not voluminous, it is important to capture the uncertainty on the sentiment prediction made by social contexts. In order to instantiate the probabilistic measure of label uncertainty we presented in this section, we use the data to set hyperparameters K_{soccer} and K_{NFL} that capture the previous knowledge on the coin that control the relationship between messages and author's groups over time. We found $K_{soccer} = 12000$ and $K_{NFL} = 6000$ as the value that maximizes the Pearson correlation that relates ΔV and $\Delta\theta(K)$ (Equation 4.3). Figure 4.7 compares, for the SOCCER dataset, the theoretical label uncertainty prediction with the empirical accuracy obtained for each volume

bin; the approximation is reasonable, and results are similar for the NFL dataset.

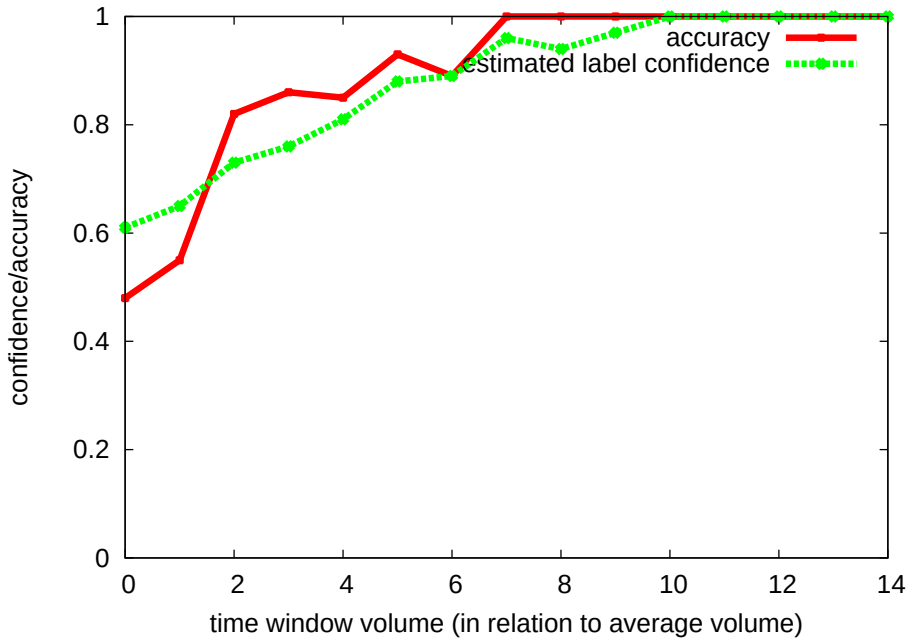


Figura 4.7. Difference between theoretical confidence estimate and empirical accuracy obtained for time windows of tweet volume = x times the average.

Figure 4.8 shows the convex shape of the Pearson correlation measure (Equation 4.3) as we increase the hyperparameter K_{soccer} in the coin model. On the red curve, we plot the absolute error between the predicted and empirical accuracy for each value of K_{soccer}, to show that the maximum of the Pearson correlations coincides with the minimum of the absolute error curve. Results are similar for the NFL dataset, and demonstrate that exploring the sequence of time-windows to smooth the measure of the coin bias θ is a simple and effective strategy.

4.3 A Feature Representation inspired by the Extreme-Average Report Imbalance

In the last section, we demonstrated the predictive power of social contexts induced by the positive-negative report imbalance and the segmentation of users into polarizing groups. In addition to the low accuracy on low-time volume windows, using just S and ignoring content D is restrictive due to two reasons:

1. Sentiment prediction does not improve over time, since knowledge from past time windows is not carried to recent time windows. Improving performance as more data is

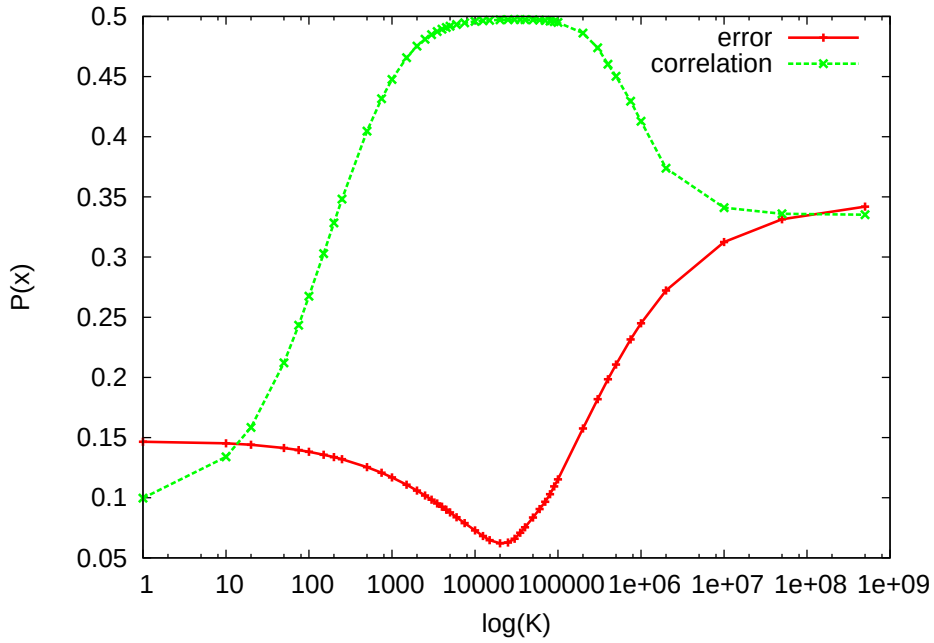


Figura 4.8. Choice of hyperparameter K_{soccer} as the value that maximizes Equation 4.3; Pearson correlation maximum coincides with the best empirical measurement of uncertainty.

processed is a basic requirement for any machine learning approach;

2. It enforces that $Y_{G_A,t} = + \rightarrow Y_{G_B,t} = -$, what is generally acceptable, given the polarized nature of polarized debate, but it is not capable of capturing more complex variations of sentiment, where members of $|G_A|$ and $|G_B|$ can share a similar sentiment at the same time, or different intensities of sentiment.

We take inspiration on the social psychology finding that describes how humans' decision on expressing their feelings is increased by the strength of the sentiment they are experiencing [7; 39; 38; 82] (which we call, for short, as **extreme-average report imbalance**) to devise a textual feature representation (and, hence, a feature selection strategy) specially designed to track sudden variations of sentiment on evolving and dynamic social streams and that makes use of the textual feature vector D_t to improve accuracy on sentiment prediction.

It is widely known that the underlying text representation impacts the performance of text mining and linguistics applications [68; 146]; different *feature definition* choices (part-of-speech features, bag-of-words, n-grams etc), *feature weighting* schemes (such as binary, TF and TF-IDF) and *feature selection* approaches can be suitable for different tasks – such as text classification, text clustering and search [146; 174]. When the textual data arrives as a stream, an adequate choice of text representation is even more critical:

- The potentially infinite size of the stream limits the storage of an ever growing high dimensional feature space, what increases the need for adequate feature representation/selection that keeps the feature space as compact as possible [80].

- Static text representations (such as TF-IDF) may not be optimized to nonstationary text streams, since they do not capture adequately the dynamic nature of the feature probability distribution [81; 68], which is strongly affected by emerging new topics and real-world events.

As explained in Section 4.2.1, D_t is the feature vector extracted from messages written during time window W_t:

$$D_t = [w_{t1}, w_{t2}, ..., w_{tM}]$$

and w_{tj} is the weight of the j-th feature in D_t. Instead of adopting traditional term frequency (TF) or term-frequency plus inverse document frequency (TF-IDF) as weights, we exploit the fact that time-windows have a varying volume of messages and, according to the extreme-average report imbalance, more people post a message when affected by an emotional, strong feeling. As a consequence, emotional content is likely to be concentrated on spikes of activity in social streams at a greater frequency than low-emotional terms. Let $\overline{W_t} = \frac{\sum_{k=0}^{t} ||W_t||}{N}$ be the average volume of messages sent in each time window up to the t-th time window and $\overline{W_{t,term}} = \frac{\sum_{k=0}^{t} ||W_k|term \in D_k||}{N_k}$ be the same measure, but considering only time windows that contain $term$. We then define $w_{t,term}$ as:

$$w_{t,term} = \frac{\overline{W_{t,term}}}{\overline{W_t}} \tag{4.5}$$

where $w_{t,term}$ measures how the occurrence of $term$ between $[W_0, W_t]$ is correlated to high-volume time windows. $w_{t,term} = 1$ means that $term$ appears on time windows whose volume are, on average, equal to the average time window volume, and thus it indicates that the term is not expected to be associated with strong emotions (e.g., spikes). A term with $w_{t,term} = 5$ means that $term$, on average, appears on time windows whose volume are five times greater than the average. We name these terms as *high-arousal* terms, since they are associated with moments where the crowd being monitored felt motivated to react and express feelings and opinions, caused by the fact that highly emotional feelings *activate* people and drive them to action [15]).

Figure 4.9 provides empirical evidence that the arousal feature space is adequate to capture sentimental n-grams by correlating the arousal measure with two features commonly associated with sentiment – the use of word lengthening [28] (as on "cooooooooooool") and the use of uppercase. The more arousal we associate with a term (n-gram), the greater is the chance it is written using one of these two linguistic indicators. In Tables 4.2 and 4.3, we display the top features in each dataset according to arousal and TF-IDF. In brackets, we show the value of arousal identified for each term; high-arousal n-grams are clearly more sentimental than TF-IDF.

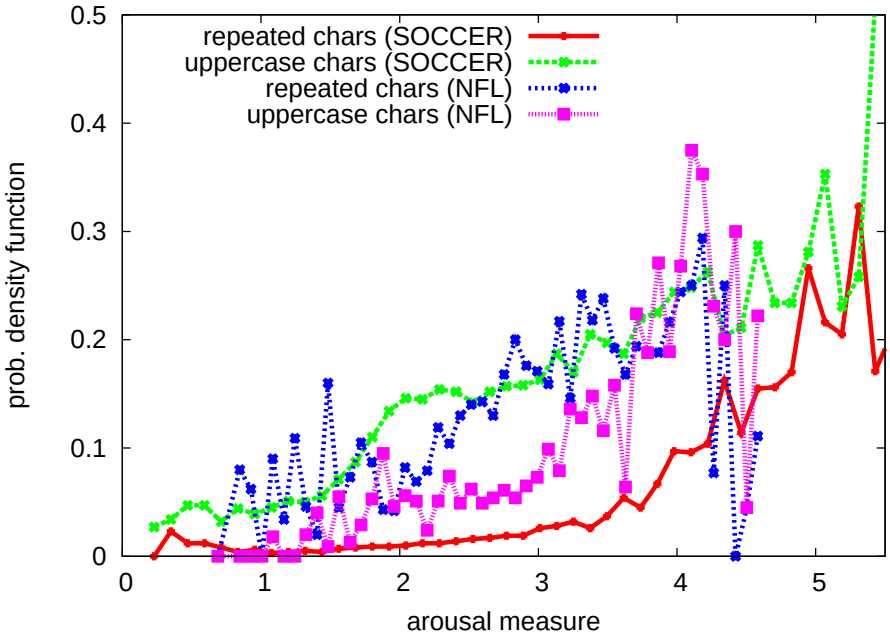


Figura 4.9. Indicators of excitement (use of uppercases and repeated characters) correlate with term arousal measure.

Tabela 4.2. Top 5 features for NFL dataset, according to *arousal* and TF-IDF representations. Arousal values are in brackets.

arousal	**TF-IDF**
PACKERS_WIN_SUPERBOWL (3.54)	yu_know_what
SUPER_BOWL_CHAMPIONS!!! (3.53)	you_would_think
YEAH! (3.43)	your_quarterback_is
superbowl_xlv_champions (2.65)	you_lost_money
touchdown!! (2.34)	you_imagine_how

High-arousal terms and concept drift. There has been significant efforts to perform effective classification on text streams under concept drift environments; the most common strategy is to employ forgetting and weighting mechanisms that decrease the importance of

Tabela 4.3. Top 5 features for SOCCER dataset, according to *arousal* and TF-IDF representations. Arousal values are in brackets.

arousal	**TF-IDF**
great_goal (7.53)	win!
goooooooooooool (6.80)	gol_from_team
he_scores(5.31)	an_equalizer
GOOOL (5.00)	go!
penalty_for_team (3.34)	he_shoots

old instances of data and force the stream classifier to focus on recent instances [184]. We follow a different strategy: instead of trying to restrict learning to recent examples, we design a dynamic feature space, where at any given time the feature space is defined by the terms selected using $arousal$ as a selection criterium. As a consequence, we are capable of quickly identifying, on spikes of activity, new features with high predictive power that may appear or gain importance over time (i.e., high values of *arousal*) that become important for sentiment classification.

When a spike occurs and (potentially) changes the dominant sentiment in the stream, due to a real world event which immediately affect users' happiness, adapting the model to such concept drift is challenging if the stream model is strongly built on past data [81]. Tackling concept drift at the feature representation stage has the advantage that unlike instance weighting and forgetting mechanisms, useful knowledge from the past is never discarded, what could harm classification perfomance [81]. In pratice, this means that we use information from old spikes to predict the sentiment at the current time window, what may be especially useful when the label is incorrectly predicted by the model we presented in Section 4.2.1.

4.3.1 Experimental Evaluation

We incorporate the textual feature vector D_t in a learning model by interpreting $P(Y|S)$ estimates from Section 4.2.1 as *probabilistic labels* (or *soft labels*), which can then be incorporated into a variety of supervised learning algorithms [142; 126]. We have chosen to employ a version of Multinomial Naive Bayes extended to consider probabilistic labels [135]. We make this choice because of the easiness to extend Naive Bayes to incorporate probabilistic labels and its suitability for stream classification, since conditional term-class probabilities can be easily updated as more data is processed. Our features correspond to unigrams, bigrams and trigrams represented by term-arousal weights.

Figure 4.10 shows how accuracy varies, in the SOCCER dataset, as we vary the number of features we include in the model, considering both our term arousal representation and

the traditional TF-IDF representation. We varied a threshold at the time window level, i.e., we included in the model the top K-ranked features on each time window. In addition to being more effective, the term arousal representation allows the sentiment model to be very compact, since the best accuracy was obtained by considering just the top 50 terms on each time window. Results are similar in the NFL dataset.

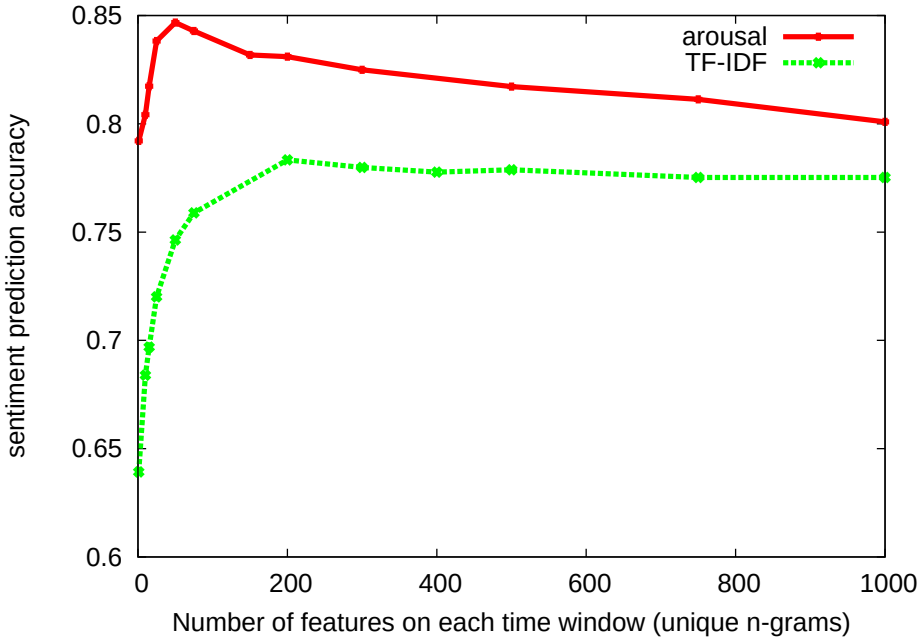


Figura 4.10. Accuracy vs top-K features comparing term-arousal and TF-IDF feature representation – SOCCER dataset.

In Figure 4.11 we show the increase on accuracy per volume bin, when adding textual features to the model. The increase on accuracy in lower-volume bins can be interpreted as the "transfer" of the reliable social context from spikes to the lower-volume time windows through the terms: when a high-arousal term is used on a low-volume time window, it contributes to the correct prediction of such time intervals.

4.3.2 Real-time sentiment tracking of live matches

To illustrate the usefulness and the utility of our combined label acquisiton/feature representation method, we now analyze the sentiment of the crowds expressed on Twitter during some interesting matches. For each match, we show the variation on the sentiment score over time in conjunction with the overall volume of tweets from each crowd. The scores are obtained by computing the ratios between the positive and negative probability estimates of the Naive Bayes classifier. Figure 4.12 shows the reactions of the supporters during SuperBowl 2011:

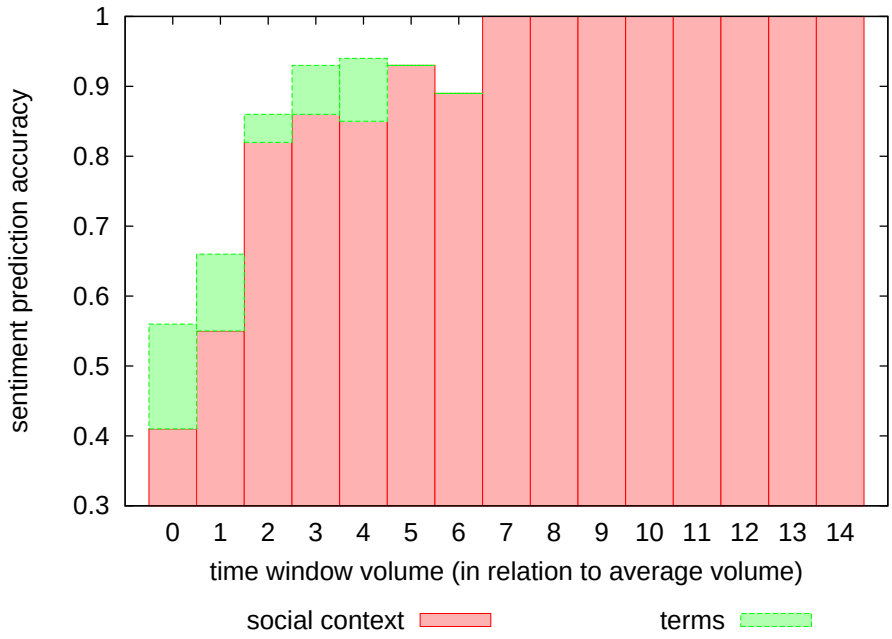


Figura 4.11. High-arousal n-grams carry the informative social contexts from the spikes to subsequent low-volume time windows – SOCCER dataset.

1. The Green Bay Packers score two touchdowns in the first quarter, reflected on the two spikes of happiness before 200'.
2. At 200' the Steelers scores a touchdown, and, after another touchdown at 240', the mood of Steelers' fans are better than Packers for a significant part of the match.
3. After a sequence of touchdowns from both teams between 320' and 350', the game comes to an end at 360' and Packers is proclaimed SuperBowl winners. Note that the majority of changes in the dominant sentiment of each crowd occur after a spike in the volume of messages, indicating that users are reacting to events. Note, also, that after the spike at 360' related to Packers' victory, our content-based classifier is capable of keeping track of the positive sentiment towards Packers, in part because of high-arousal terms such as those shown in Table 4.2.

In the 2012 SuperBowl, played on February 5th, we also detected changes in crowd's humour, as shown in Figure 4.13:

1. The New York Giants started the game scoring 2-0 at 158' and 9-0 with a touchdown at 168'.
2. The Patriots scored two touchdowns in a row, at 224' and 265', reversing the expectations about the game outcome.

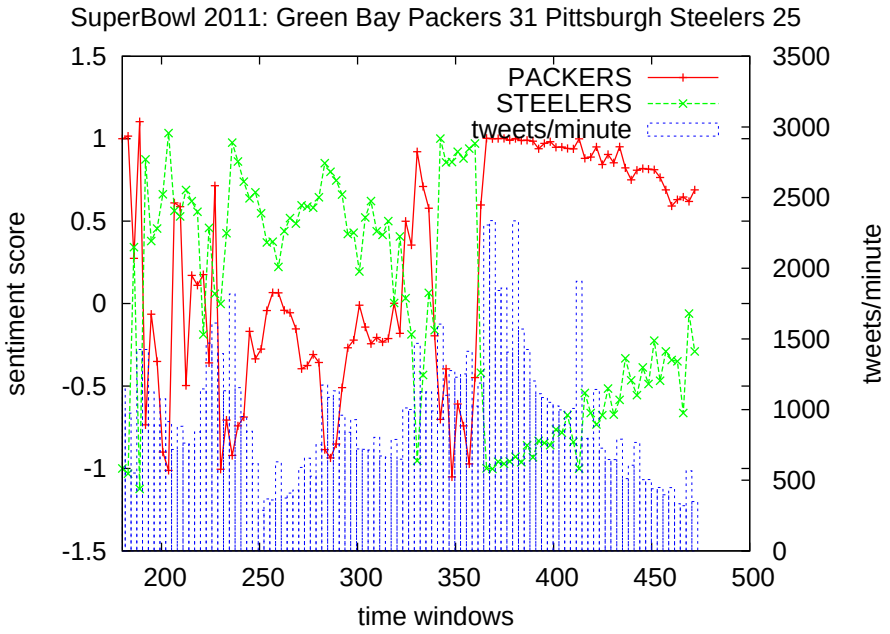


Figura 4.12. Sentiment variation during SuperBowl 2011 – Packers vs Steelers.

3. The Giants managed to score a touchdown in the last minute of the game and were proclaimed the 2012 SuperBowl champions at 298', generating a long period of happiness on their supporters, whereas Patriots supporters were upset.

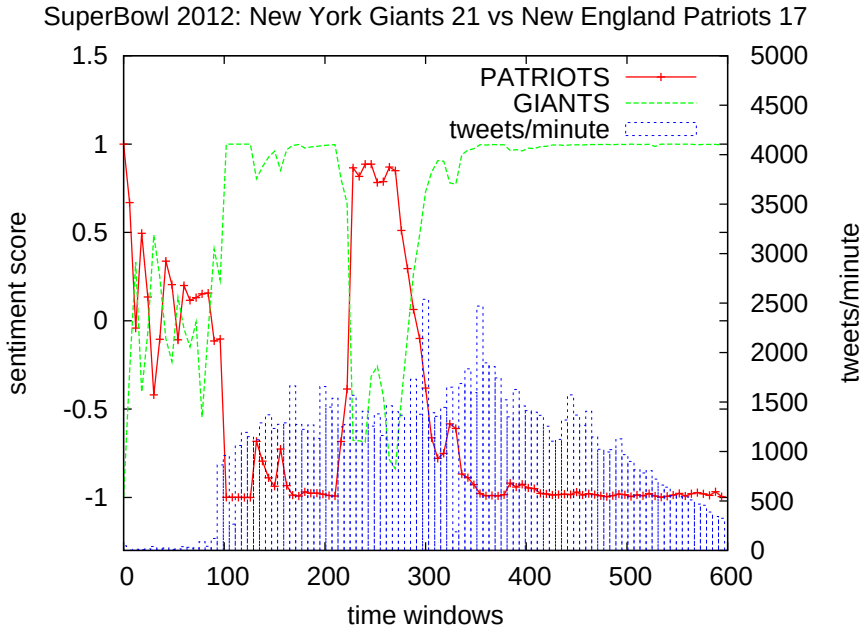


Figura 4.13. Sentiment variation during SuperBowl 2012 – Giants vs Patriots.

Soccer. We also illustrate our results with two matches of the last round of the 2011 Brazilian Soccer League. Although games last for 90 minutes, we also show crowd sentiment before and after the match duration. In Figure 4.14, team Cruzeiro comfortably beats his fierce rival Atletico by a surprising score of 6-1, scoring two goals in the early minutes of the match. Our model was able to correctly capture the positive reactions of Cruzeiro fans, and negative reactions of Atletico supporters. The second match, in Figure 4.15, showed a totally different pattern: Vasco and Flamengo played at the last round of the Brazilian 2011 Soccer League and Vasco needed to win in order to have any chance of winning the championship title:

1. At 149', Vasco scored, and our algorithm detected a sudden burst of positive sentiments for Vasco and negative sentiments for Flamengo.
2. At minute 199', however, Flamengo scored (note the spike in volume of tweets), vanishing any chances of Vasco winning the title. Our algorithm detected a sharp negative spike for Vasco in that moment. Even after conceding a goal, Vasco supporters were still upset, as expected; this illustrates the capacity of our algorithm in learning from spikes and using the learned term polarities on the subsequent time intervals.
3. Note that we have been able to track different supporters' reactions, even during "similar" events: although Atletico scored against Cruzeiro at 220', it was already losing by 5-0, what kept Cruzeiro supporters at a better mood. On the other hand, Flamengo's tie goal against Vasco was a much more important one, and, even though Vasco was not losing the game, that goal vanished their chances of winning the title.

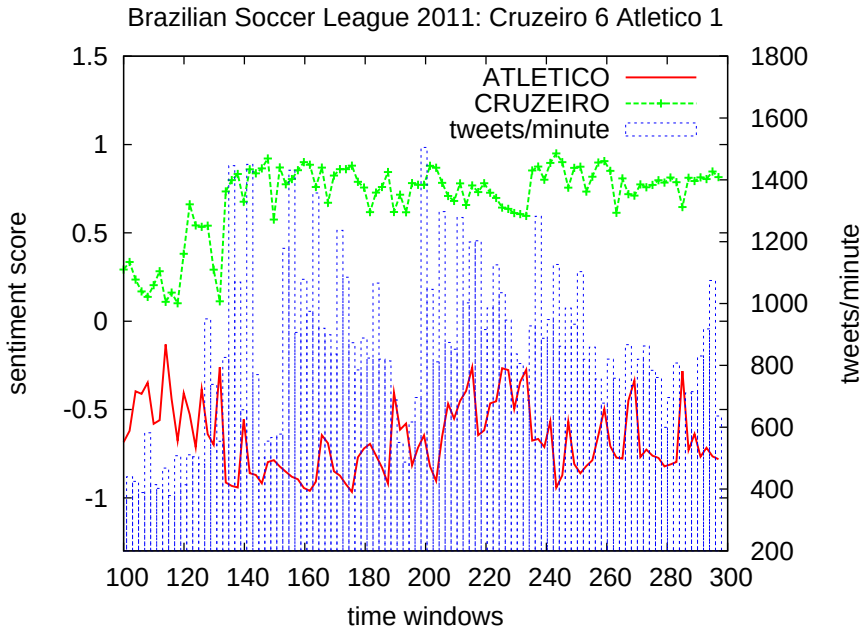


Figura 4.14. Sentiment variation during Brazilian Soccer League match – Cruzeiro vs. Atletico.

4.4 Related Work on Self-Report Imbalances and Sentiment Analysis

Social media data has been successfully used to detect real-world events such as disease outbreaks [34], earthquakes [139] and recurring events such as goals and touchdowns in sports matches [93]. Most of these researches are not focused on the deviation between self-reported data and real data; it is implicitly assumed that the number of users who decide to react and comment on the events being monitored will be large enough to allow detection. However, the self-reported nature of social media can strongly impact the observed social data, as observed by [82]: if we search in Twitter for the words “breathing” and “drinking water”, we may end up (wrongly) concluding that people usually drink more water than breath in their daily lifes. Some recent works try to compensate these biases in analysis of political debate, by observing that a small fraction of people intensively self-report their political opinions, while a silent majority does not [123], what can dramatically change conclusions and statistics on political behavior. Differently from these works, we stress that we aim to use self-reporting bias and the social/temporal contexts it creates to our benefit, in the design of better opinion analysis models, rather than correcting its effects.

Our work is closely related to research that exploits opinion holder biases’ to perform sentiment analysis. Especially un the political domain, it is known that biases on opinion

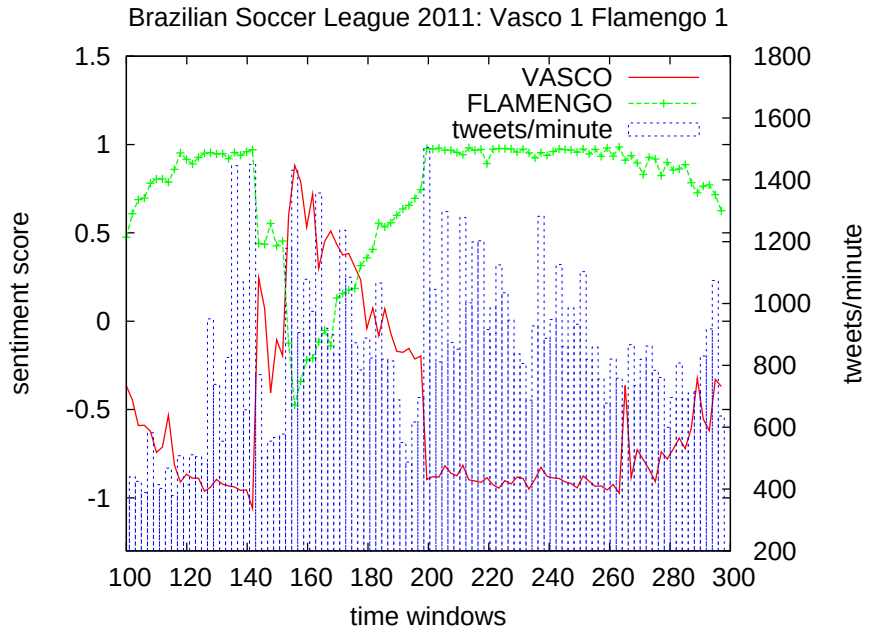


Figura 4.15. Sentiment variation during Brazilian Soccer League - Vasco vs. Flamengo.

holders highly correlate to the type of opinion they express, and that social contexts based on groups of people with similar viewpoints provide useful signals for opinion analysis [29; 53; 104]. We add to these group-based social contexts a temporal perspective to explore the correlation between the real-world events taking place and the users currently reacting to what they are observing. To the best of our knowledge, this is the first attempt to detect positive and negative sentiment expressed on online media by capitalizing on the reasons that stimulate people to communicate more or less their feelings.

Sentiment analysis is still focused on static scenarios such as product reviews [131], on which lexicons of positive and negative words and traditional supervised machine learning techniques have been quite successful [155]. We are interested in sentiment analysis as a stream data mining task, a setting which requires learning algorithms to constantly update and refine data description models, in face of the time-changing characteristics of the data [52; 21]. The simultaneous presence of concept drift and lack of labeled data makes real-time sentiment analysis an even harder problem, since some standard solutions from one challenge make assumptions that do not hold in the other. The state-of-the art solution for coping with the scarcity of labeled data, *semi-supervised learning*, makes use of both labeled and unlabeled data for model generationand has also been applied to sentiment analysis [106]. However, due to the nonstationary characteristic of social streams, the usefulness of a few initially available labeled examples may be limited since they can become quickly outdated [44]. Conversely, the traditional approach for dealing with concept drift

on nonstationary data is to incrementally update the model through fresh, recently-acquired labels that are provided by the stream [167], but this solution may not be feasible due to the lack of labeled data. In terms of machine learning approaches, our algorithm is best related to *distant supervision* [60], which generates labeled data not by manual inspection of individual instances, but by applying some sort of heuristic/rule which outputs noisy labels. While distant supervision has considered emoticons as the source of labels, we take inspiration on social psychology patterns that guide people's reactions.

4.5 Wrap Up

Real-time sentiment analysis is a difficult task; labeled data is usually not available to support supervised classifiers, and debate about monitored topics may turn into unpredictable discussions. We propose solutions to these challenges based on the different propensity users have on disclosing positive and extreme feelings, in comparison to negative and average feelings.

Since we mapped the usage of the social information on two machine-learning sub-tasks – acquisition of labeled data and feature representation – our work is orthogonal to current and future supervised models for real-time sentiment analysis. Depending on the characteristics of the domain and the social media platform, one or other sub-task may benefit more from our models.

One future direction is to better investigate the impact of time window sizes. In addition to automatically determine the optimal window size (or make it dynamic), analyzing effects of different window sizes in our models may unveil new patterns on how social media users react to real-world events.

Capítulo 5

Conclusions and Final Remarks

In this dissertation, we establish connections between well-known social psychology theories and machine learning algorithms that process social streams containing opinions regarding polarized topics. We map these theories, which are related to how opinions holders' opinions are predictably biased, into new signals that are employed by machine learning models to classify and organize content according to the sentiment and opinion it conveys regarding entities of interest (such as candidates, celebrities and organizations). Our dissertation sustains the hypothesis that, if a topic debate is recognizably polarized, a simple and strong hypothesis hold: **opinion holders** and **opinions** are not independent, but **correlated**. We make the following contributions:

1. First, we *measure* the strength of polarization on (online) social networks with respect to a given topic discussion. We demonstrate that the current network science metric widely used to measure polarization (modularity) is not well suited to discriminate between polarization and absence of polarization; we then propose and evaluate two additional metrics based on the social network structure that, as we will demonstrate, captures more accurately the social phenomena of polarization.

2. Second, we propose new methods for processing and interpreting opinions expressed on online polarized debates by uncovering from the social psychology literature well-established social theories that describe how people form their opinions on polarized discussions. We use these theories as foundations for new signals that enable sentiment analysis method to operate on polarized discussions that arrive in the form of *social streams* – an **evolving**, **bursty** and **time-changing** flow of opinions.

   We believe our work is in consonance with the widely reported observation that, in practical machine learning problems, feature engineering tends to be one of the tasks that yields better improvement on machine learning models [168]. Instead of going

in the direction of more complex models to perform sentiment analysis (that use, for instance, deep learning [145]), we prefer to follow the path that, once the right attributes are known, simple and faster models can be adopted [168]. Moreover, the signals we propose are strongly backed up by previous research on empirical and theoretical social sciences and social psychology.

3. Our third contribution is related to the observation that, on many domains, we have **more than two viewpoints** in conflict with respect to a topic, as on multipartisan political systems. Differently from the classical case of bipolarization of opinions, in multipolarized social networks we observe more complex relationships among sides, rather than the duality support/antagonism. In addition to highlighting inconsistencies that are hidden on bipolarized network analyses, we propose a page-rank based model that infers antagonism relationships among social communities in such a setting. Instead of relying only on positive seeds, we find negative seeds by exploiting a implicit negative feedback assumption that opinion holders' do not react to messages that are contrary to their viewpoints in the same intensity they do when messages endorse their current opinions.

We believe our work also contribute to social sciences in the sense that we validate and observe empirical and theoretical findings in their field in additional domains in the online world, such as the preference for disclosing positive and extreme feelings in Twitter during polarized events.

As a growing fraction of web content is generated in the form of social streams, we believe there is a promising opportunity to build rich applications that track the emotional reactions of social media users during dynamically changing and potentially polarizing events such as sports matches, political debates and live breaking news. Traditional sentiment analysis, however, is not designed to operate on the stream setting, since the field has focused its attention on extracting opinions from static text such as product and movie reviews. We believe that our work can be helpful in that direction.

We also shed light on the fact that using social media platforms as a tool to infer the public opinion should be taken with caution, due to the high bias on the opinions expressed by those (not only by humans, but also by automated bots) who decided to give an opinion publicly.

5.1 Publications

Here we list the main publications that are associated with this dissertation:

1. Pedro Calais Guerra, Adriano Veloso, Wagner Meira Jr; Virgilio Almeida. From Bias to Opinion: **A Transfer-Learning Approach o Real-Time Sentiment Analysis**. 17h International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD), 2011, San Diego, CA. Proceedings of the 17h International Conference on Knowledge Discovery and Data Mining, 2011.

2. Pedro Calais Guerra, Loic Cerf, Thiago Porto, A. Veloso, Wagner Meira Jr., Virgilio Almeida. **Exploiting Temporal Locality to Determine User Bias in Microblogging Platforms**. Journal of Information and Data Management (JIDM), v.2, p. 273-288, 2011.

3. Pedro Calais Guerra, Wagner Meira Jr., Claire Cardie, Robert Kleinberg. **A Measure of Polarization on Social Media Networks based on Community Boundaries**. 7th International AAAI Conference on Weblogs and Social Media (AAAI ICWSM), 2013, Boston, USA.

4. Pedro Calais Guerra, Wagner Meira Jr., Claire Cardie. **Sentiment Analysis on Evolving Social Streams: How Self-Report Imbalances Can Help**. 7th International ACM Conference on Web Search and Data Mining (ACM WSDM), 2014, New York City, USA.

5.2 Next Steps

We envisage a series of future research directions.

5.2.1 Characterizing and Modeling Self-Reporting Bias at User-Level

We plan to enrich the social context we use to track sentiments by exploring the reaction patterns not only at group-level, but at user-level and on multi-group levels. At the user level, we can uncover different, more complex behavior of social media user posting patterns. Are there users which, in contrast to the dominant pattern, prefer to comment on negative experiences for their opposing sides than on positive events of their own side? At multi-group level, we may exploit the different relationships between polarized groups to generate more informative social contexts. For instance, supporters from rival teams are likely to follow and react whenever their rivals are being defeated, and that information could be embedded in the social context.

5.2.2 New Opinion Mining Tasks

So far, opinion mining and sentiment analysis have focused on, given an input $< oh, d, t >$, predict the polarity $p = \{+, -\}$ of the document d written by opinion holder oh regarding target entity t. We envisage as interesting fields of research the concretization of other common data mining tasks, other than classification, that could generate new, relevant and previosly unknown patterns related to opinions and opinion holders:

1. **Finding rare opinions:** Some topics of discussion seem to have a dominant opinion, either positive or negative. For example, the vast majority of Brazilians think that the "Mensalao" scandal really happened and all politicians involved are guilt. An interesting opinion mining task is then to find rare opinions: are there opinion holders who have an opinion which is totally different from the dominant one?

2. **Finding surprising opinions:** In some situations either positive or negative opinions are common, but a instance $< oh, t, \{+, -\} >$ is not expected. For example, a partisan supporter of a candidate showing criticism to his own candidate is not an expected opinion. To find unexpected $< oh, t, \{+, -\} >$ tuples it a task of interest in our research. It can unveil dense, polemic and interesting opinions, since it motivated an opinion holder to give an opinion which is contrary to his bias.

3. **Opinion Entropy:** The sentiment analysis task on polarized debate also brings new questions that we do not witness on the "classical" sentiment analysis product-review scenario: given the high-biased nature of opinions, what is the "value" of a biased opinion? In some sense, if everybody issues opinions which match their expected bias (either supporting their favorite side or criticizing and opponent), the overall "opinion entropy" of the system is zero, i.e., we do not learn too much from the opinions, because they only reflect people's bias. In practice, we can understand an opinion as a sum of two factors: a combination of what the person has seen on the past regarding that issue/topic and the fact currently being analyzed. On high-biased people, their opinions reflect much more what they already think on the subject than an analysis of the current facts. How to "unbias" the public opinion is an interesting reseach question here, because bias, in some aspect, insert "noise" in measuring the public reaction to events.

   Still in this direction, we observe that for a set of users U and a set of opinions O, different pairwise combinations of pairs $(u \in U, o \in O)$ can represent opinions which are semantically different on the aggregate, by generating different levels of "opinion entropy". The interesting problem here is to propose a measure of "opinion entropy" and detect unexpected, interesting (and perhaps more sincere?) opinions.

4. **Identifying Tipping Points in Public Opinion:** Finding changes in public opinion is useful for a number of reasons, allowing marketers to react to a tipping point in people's thoughts in reaction to real-life events. Although there are some research on that direction, content analysis is the dominant approach [5]. By observing the evolution of the social graph over time, we may be able to detect significant changes in its structure which, ultimately, represent changes in people's bias and viewpoints.

Referências Bibliográficas

[1] Abramowitz, A. & Saunders, K. (2005). Why cant we all just get along? the reality of a polarized america. *The Forum: A Journal of Applied Research in Contemporary Politics*, 2.

[2] Abu-Jbara, A.; Ezra, J. & Radev, D. R. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. Em *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 596--606.

[3] Adamic, L. A. & Glance, N. (2005). The political blogosphere and the 2004 u.s. election: divided they blog. Em *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pp. 36--43, New York, NY, USA. ACM.

[4] Agrawal, R.; Rajagopalan, S.; Srikant, R. & Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. Em *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pp. 529--535, New York, NY, USA. ACM.

[5] Akcora, C. G.; Bayir, M. A.; Demirbas, M. & Ferhatosmanoglu, H. (2010). Identifying breakpoints in public opinion. Em *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 62--66, New York, NY, USA. ACM.

[6] Andersen, R.; Chung, F. & Lang, K. (2006). Local graph partitioning using pagerank vectors. Em *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pp. 475--486, Washington, DC, USA. IEEE Computer Society.

[7] Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1):5--17.

[8] Andreoni, J. & Mylovanov, T. (2012). Diverging opinions. *American Economic Journal: Microeconomics*, 4.

[9] Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. Em *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pp. 81--87, Stroudsburg, PA, USA. Association for Computational Linguistics.

[10] Awadallah, R.; Ramanath, M. & Weikum, G. (2012). Opinions network for politically controversial topics. Em *Proceedings of the First Edition Workshop on Politics, Elections and Data*, PLEAD '12, pp. 15--22, New York, NY, USA. ACM.

[11] Bakshy, E.; Messing, S. & Adamic, L. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*.

[12] Balasubramanyan, R.; Cohen, W. W.; Pierce, D. & Redlawsk, D. P. (2012). Modeling polarizing topics: When do different political communities respond differently to the same news? Em *ICWSM*. The AAAI Press.

[13] Baron, J. (2006). *Thinking and Deciding*. Cambridge University Press.

[14] Barrett, L. F. & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality & Social Psychology*, 74(4):967--984.

[15] Barrett, L. F. & Russell, J. A. (1999). The Structure of Current Affect: Controversies and Emerging Consensus. *Current Directions in Psychological Science*, 8(1):10--14.

[16] Bazarova, N. N.; Choi, Y. H.; Schwanda Sosik, V.; Cosley, D. & Whitlock, J. (2015). Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. Em *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pp. 154--164, New York, NY, USA. ACM.

[17] Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I. & Culotta, A., editores (2009). *Bayesian Belief Polarization*. Curran Associates, Inc.

[18] Berger, J. (2013). *Contagious: Why Things Catch On*. Simon & Schuster.

[19] Bermingham, A. & Smeaton, A. F. (2010a). Classifying sentiment in microblogs: is brevity an advantage? Em *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 1833--1836, New York, NY, USA. ACM.

[20] Bermingham, A. & Smeaton, A. F. (2010b). Crowdsourced real-world sensing: sentiment analysis and the real-time web. Em *AICS 2010 - Sentiment Analysis Workshop at Artificial Intelligence and Cognitive Science*.

[21] Bifet, A. & Kirkby, R. (2009). Data stream mining: a practical approach.

[22] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[23] Blendon; Young & Hemenway (1996). The american public and the gun control debate. *JAMA*, 275(22):1719–1722.

[24] Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of community hierarchies in large networks. *CoRR*, abs/0803.0476.

[25] Bond, R.; Faris, C.; Jones, J.; Kramer, A.; Marlow, C.; Settle, J. & Fowler, J. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295--298.

[26] Boyd, D.; Golder, S. & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. Em *Proceedings of the 43rd Hawaii International Conference on Social Systems (HICSS)*. IEEE.

[27] Brew, A.; Greene, D. & Cunningham, P. (2010). Using crowdsourcing and active learning to track sentiment in online media. Em *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pp. 145--150, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[28] Brody, S. & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. Em *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA. ACL.

[29] Calais, P. H.; Veloso, A.; Meira, Jr, W. & Almeida, V. (2011). From bias to opinion: A transfer-learning approach to real-time sentiment analysis. Em *Proc. of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, CA.

[30] Chapelle, O.; Schölkopf, B. & Zien, A. (2006). Semi-supervised learning. MIT Press.

[31] Chen, L.; Weber, I. & Okulicz-Kozaryn, A. (2014). U.S. religious landscape on twitter. Em *Social Informatics - 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, pp. 544--560.

[32] Choi, Y.; Kim, Y. & Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation. Em *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, pp. 37--44, New York, NY, USA. ACM.

[33] Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Flammini, A. & Menczer, F. (2011). Political polarization on twitter. Em *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[34] Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. Em *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 115--122, New York, NY, USA. ACM.

[35] Dandekar, P.; Goel, A. & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796.

[36] Dave, K.; Lawrence, S. & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Em *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pp. 519--528, New York, NY, USA. ACM.

[37] De Choudhury, M.; Mason, W. A.; Hofman, J. M. & Watts, D. J. (2010). Inferring relevant social networks from interpersonal communication. Em *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 301--310, New York, NY, USA. ACM.

[38] Dellarocas, C. & Narayan, R. (2006). A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth. *Statistical Science*, 21(2):277--285.

[39] Dellarocas, C. & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Manage. Sci.*, 54(3):460--476.

[40] Desrosiers, C. & Karypis, G. (2009). Within-network classification using local structure similarity. Em *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I*, pp. 260--275.

[41] Diakopoulos, N. A. & Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. Em *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pp. 1195--1198, New York, NY, USA. ACM.

[42] Diener, E.; Emmons, R.; Larsen, R. & Griffin, S. (1985). The satisfaction with life scale. *J Pers Assess*, 49(1):71–5.

[43] Dixit, A. K. & Weibull, J. W. (2007). Political polarization. *Proceedings of the National Academy of Sciences*, 104(18):7351--7356.

[44] Dyer, K. B. & Polikar, R. (2012). Semi-supervised learning in initially labeled non-stationary environments with gradual drift. Em *IJCNN*, pp. 1–9. IEEE.

[45] Earl, J.; Martin, A.; McCarthy, J. D. & Soule, S. A. (2004). The use of newspaper data in the study of collective action. volume 30, pp. 65--80. Annual Review of Sociology.

[46] Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

[47] Fortunato, S. (2009). Community detection in graphs. *CoRR*, abs/0906.0612.

[48] Fortunato, S. & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36--41.

[49] Fouss, F.; Pirotte, A.; Renders, J.-M. & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. on Knowl. and Data Eng.*, 19:355--369.

[50] Friedkin, N. E. (1998). *A Structural Theory of Social Influence (Structural Analysis in the Social Sciences)*. Cambridge University Press.

[51] Gallagher, B.; Tong, H.; Eliassi-Rad, T. & Faloutsos, C. (2008). Using ghost edges for classification in sparsely labeled networks. Em *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, New York, NY, USA. ACM.

[52] Gama, J. a.; Sebastião, R. & Rodrigues, P. P. (2009). Issues in evaluation of stream learning algorithms. Em *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pp. 329--338, New York, NY, USA. ACM.

[53] Gamon, M.; Basu, S.; Belenko, D.; Fisher, D.; Hurst, M. & König, A. C. (2008). Blews: Using blogs to provide context for news articles. Em *In Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[54] Gayo-Avello, D.; Metaxas, P. T. & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. Em *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain*.

[55] Gelman, A.; Carlin, J.; Stern, H. & Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

[56] Gerani, S.; Carman, M. J. & Crestani, F. (2010). Proximity-based opinion retrieval. Em *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'10, pp. 403--410, New York, NY, USA. ACM.

[57] Giatsoglou, M.; Chatzakou, D.; Shah, N.; Faloutsos, C. & Vakali, A. (2015). Retweeting activity on twitter: Signs of deception. Em *Advances in Knowledge Discovery and Data Mining - 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part I*, pp. 122--134.

[58] Giraud-Carrier, C. (2000). A note on the utility of incremental learning. *AI COMMUNICATIONS*, 13:215--223.

[59] Girvan, M. & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821--7826.

[60] Go, A.; Bhayani, R. & Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford.

[61] Golbeck, J. & Hansen, D. (2011). Computing political preference among twitter followers. Em *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pp. 1105--1108, New York, NY, USA. ACM.

[62] Good, B. H.; de Montjoye, Y. A. & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106+.

[63] Graham, P. (2008). How to disagree. http://www.paulgraham.com/disagree.html.

[64] Guerra, P. H. C.; Jr., W. M.; Cardie, C. & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. Em *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, Boston,MA.

[65] Gupta, P.; Goel, A.; Lin, J.; Sharma, A.; Wang, D. & Zadeh, R. (2013). Wtf: The who to follow service at twitter. Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 505--514, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[66] Hassan, A.; Abu-Jbara, A. & Radev, D. (2012a). Detecting subgroups in online discussions by modeling positive and negative relations among participants. Em *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning, EMNLP-CoNLL '12, pp. 59--70, Stroudsburg, PA, USA. Association for Computational Linguistics.

[67] Hassan, A.; Abu-Jbara, A. & Radev, D. (2012b). Extracting signed social networks from text. Em *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, TextGraphs-7 '12, pp. 6--14, Stroudsburg, PA, USA. Association for Computational Linguistics.

[68] He, Q.; Chang, K.; Lim, E.-P. & Zhang, J. (2007). Bursty feature representation for clustering text streams. Em *SDM*. SIAM.

[69] Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons, New York.

[70] Hu, N.; Zhang, J. & Pavlou, P. A. (2009). Overcoming the j-shaped distribution of product reviews. *Commun. ACM*, 52(10):144--147.

[71] Hu, X.; Tang, L.; Tang, J. & Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. Em *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13.

[72] Hu, Y.; Koren, Y. & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. Em *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 263--272, Washington, DC, USA. IEEE Computer Society.

[73] Hunter, J. (1992). *Culture Wars: The Struggle to Define America*. Sociology. Religion. BasicBooks.

[74] Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50(6):1141--1151.

[75] Jansen, B. J.; Zhang, M.; Sobel, K. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60:2169--2188.

[76] Jin, W.; Ho, H. H. & Srihari, R. K. (2009). Opinionminer: a novel machine learning system for web opinion mining and extraction. Em *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, New York, NY, USA. ACM.

[77] Jo, Y. & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. Em *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pp. 815--824, New York, NY, USA. ACM.

[78] Jordan, A. H.; Monin, B.; Dweck, C. S.; Lovett, B. J.; John, O. P. & Gross, J. J. (2010). Misery Has More Company Than People Think: Underestimating the Prevalence of Others' Negative Emotions. *Personality and Social Psychology Bulletin*, 37(1):120--135.

[79] Kannan, R.; Vempala, S. & Vetta, A. (2004). On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497--515.

[80] Katakis, I.; Tsoumakas, G. & Vlahavas, I. (2005). On the utility of incremental feature selection for the classification of textual data streams. Em *10th Panhellenic Conference on Informatics (PCI 2005)*. Springer-Verlag.

[81] Katakis, I.; Tsoumakas, G. & Vlahavas, I. (2006). Dynamic feature space and incremental feature selection for the classification of textual data streams. Em *ECML/PKDD-2006 International Workshop on Knowledge Discovery from Data Streams. 2006*. Springer Verlag.

[82] Kiciman, E. (2012). Omg, i have to tweet that! a study of factors that influence tweet rates.

[83] Kida, T. (2006). *Don't Believe Everything You Think: The 6 Basic Mistakes We Make in Thinking*. Prometheus Books.

[84] Kienpointner, M. & Kindt, W. (1997). On the problem of bias in political argumentation : an investigation into discussions about political asylum in germany and austria. *Journal of Pragmatics*, 5(27):555--585.

[85] Krackhardt, D. & Stern, R. N. (1988). Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, 51(2):123--140.

[86] Krebs, V. (2008). New political patterns.

[87] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480--498.

[88] Kunegis, J.; Lommatzsch, A. & Bauckhage, C. (2009). The slashdot zoo: mining a social network with negative edges. Em *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pp. 741--750.

[89] Kunegis, J.; Preusse, J. & Schwagereit, F. (2013a). What is the added value of negative links in online social networks? Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 727--736, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[90] Kunegis, J.; Preusse, J. & Schwagereit, F. (2013b). What is the added value of negative links in online social networks? Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 727--736, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[91] Kunegis, J.; Schmidt, S.; Lommatzsch, A.; Lerner, J.; Luca, E. W. D. & Albayrak, S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. Em *SDM*, pp. 559–. SIAM.

[92] Kushin, M. J. & Kitchener, K. (2009). Getting political on social network sites: Exploring online political discourse on facebook. *First Monday*, 14(11).

[93] Lanagan, J. & Smeaton, A. F. (2011). Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, pp. 542--545.

[94] Larson, R.; Csikszentmihalyi, M. & Graef, R. (1982). Time alone in daily experience: Loneliness or renewal? *Loneliness: A sourcebook of current theory, research and therapy*.

[95] Lazer, B. D. (2015). The rise of the social algorithm. *Science*.

[96] Leskovec, J.; Backstrom, L. & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. Em *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pp. 497--506, New York, NY, USA. ACM.

[97] Leskovec, J.; Huttenlocher, D. & Kleinberg, J. (2010a). Predicting positive and negative links in online social networks. Em *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 641--650, New York, NY, USA. ACM.

[98] Leskovec, J.; Huttenlocher, D. & Kleinberg, J. (2010b). Signed networks in social media. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 1361--1370, New York, NY, USA. ACM.

[99] Levinson, S. (1983). *Pragmatics*. Cambridge Textbooks In Linguistics. Cambridge University Press.

[100] Liben-Nowell, D. & Kleinberg, J. (2003). The link prediction problem for social networks. Em *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pp. 556--559, New York, NY, USA. ACM.

[101] Lin, C. H.; Kamar, E. & Horvitz, E. (2014). Signals in the silence: Models of implicit feedback in a recommendation system for crowdsourcing. Em *Proceedings of the*

Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada., pp. 908--915.

[102] Lin, F. & Cohen, W. W. (2010). Semi-supervised classification of network data using very few labels. Em *ASONAM*, pp. 192–199.

[103] Lin, Y.-R.; Margolin, D.; Keegan, B. & Lazer, D. (2013a). Voices of victory: A computational focus group framework for tracking opinion shift in real time. Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 737--748, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[104] Lin, Y.-R.; Margolin, D.; Keegan, B. & Lazer, D. (2013b). Voices of victory: a computational focus group framework for tracking opinion shift in real time. Em *Proceedings of the 22nd int'l conference on World Wide Web*, WWW '13.

[105] Liu, B. (2010). Sentiment analysis: A multi-faceted problem. Em *IEEE Intelligent Systems*.

[106] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis digital library of engineering and computer science. Morgan & Claypool.

[107] Livne, A.; Simmons, M. P.; Adar, E. & Adamic, L. A. (2011). The party is over here: Structure and content in the 2010 election. Em Adamic, L. A.; Baeza-Yates, R. A. & Counts, S., editores, *ICWSM*. The AAAI Press.

[108] Lo, D.; Surian, D.; Zhang, K. & Lim, E.-P. (2011). Mining direct antagonistic communities in explicit trust networks. Em *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pp. 1013--1018, New York, NY, USA. ACM.

[109] Lord, C.; Ross, L. & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098--2109.

[110] Macskassy, S. A. & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935--983.

[111] Masud, M. M.; Woolam, C.; Gao, J.; Khan, L.; Han, J.; Hamlen, K. W. & Oza, N. C. (2011). Facing the reality of data stream classification: coping with scarcity of labeled data. *Knowl. Inf. Syst.*, 33(1):213–244.

[112] McCright, A. M. & Dunlap, R. E. (2011). The politicization of climate change and polarization in the american public's views of global warming, 2001-2010. *The Sociological Quarterly*.

[113] McPherson, M.; Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

[114] McRaney, D. (2012). *You Are Not So Smart: Why You Have Too Many Friends on Facebook, Why Your Memory Is Mostly Fiction, and 46 Other Ways You're Deluding Yourself*. Gotham Books.

[115] Melville, P.; Gryc, W. & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. Em *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, New York, NY, USA. ACM.

[116] Meshi, D.; Morawetz, C. & Heekeren, H. R. (2013). Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in Human Neuroscience*, 7(439).

[Metaxas et al.] Metaxas, P. T.; Mustafaraj, E. & Gayo-Avello, D. How (not) to predict elections. Em *2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011*.

[118] Mey, J. (2001). *Pragmatics: An Introduction*. John Wiley & Sons.

[119] Milyo, J. & Groseclose, T. (2005). A measure of media bias. Working Papers 0501, Department of Economics, University of Missouri.

[120] Mitchell, A. & Hitlin, P. (2013). Twitter reaction to events often at odds with overall public opinion. Pew Research Center. http://www.pewresearch.org/2013/03/04/\discretionary{-}{}{}twitter-re

[121] Mouw, T. & Sobel, M. (2001). Culture wars and opinion polarization: The case of abortion. *American Journal of Sociology*.

[122] Mowbray, M. (2010). The twittering machine. Em Filipe, J. & Cordeiro, J., editores, *WEBIST (2)*, pp. 299–304. INSTICC Press.

[123] Mustafaraj, E.; Finn, S.; Whitlock, C. & Metaxas, P. T. (2011). Vocal minority versus silent majority: Discovering the opinions of the long tail. Em *SocialCom/PASSAT*.

[124] Mustafaraj, E. & Metaxas, P. T. (2011). What edited retweets reveal about online political discourse. Em *Analyzing Microtext*, volume WS-11-05 of *AAAI Workshops*. AAAI.

[125] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577--8582.

[126] Nguyen, Q.; Valizadegan, H. & Hauskrecht, M. (2011). Learning classification with auxiliary probabilistic information. Em *Proc. of the 11th IEEE Int'l Conf. on Data Mining*, ICDM '11, Washington, DC, USA. IEEE Computer Society.

[127] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175--220.

[128] O'Connor, B.; Balasubramanyan, R.; Routledge, B. & Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. Em *Proceedings of the Int'l AAAI Conference on Weblogs and Social Media*.

[129] Palla, G.; Derényi, I.; Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814--818.

[130] Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

[131] Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1--135.

[132] Paul DiMaggio, John Evans, B. B. (1996). Have American's Social Attitudes Become More Polarized? *American Journal of Sociology*, 102(3):690--755.

[133] Plous, S. (1993). *The psychology of judgment and decision making*. McGraw-Hill, New York.

[134] Pohl, R. (2005). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Taylor & Francis.

[135] Ramakrishnan, G.; Chitrapura, K. P.; Krishnapuram, R. & Bhattacharyya, P. (2005). A model for handling approximate, noisy or incomplete labeling in text classification. Em *Proceedings of the 22nd international conference on Machine learning*, ICML '05, New York, NY, USA. ACM.

[136] Redlawsk, D. P.; Civettini, A. J. W. & Emmerson, K. M. (2010). The affective tipping point: Do motivated reasoners ever "get it"? *Political Psychology*, 31:563--593.

[137] Rost, M.; Barkhuus, L.; Cramer, H. & Brown, B. (2013). Representation and communication: challenges in interpreting large social media datasets. Em *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, New York, NY, USA. ACM.

[138] Ryan, T. & Xenos, S. (2011). Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 27(5):1658 – 1664.

[139] Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. Em *Proceedings of the 19th international conference on World wide web*, WWW '10, New York, NY, USA. ACM.

[140] Samangooei, S.; Preotiuc-Pietro, D.; Cohn, T.; Niranjan, M. & Gibbins, N. (2012). Trendminer: An architecture for real time analysis of social media text. *AAAI Publications, Sixth International AAAI Conference on Weblogs and Social Media*.

[141] Schafer, J. B.; Konstan, J. & Riedl, J. (1999). Recommender systems in e-commerce. Em *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, pp. 158--166, New York, NY, USA. ACM.

[142] Sheng, V. S.; Provost, F. & Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. Em *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, New York, NY, USA. ACM.

[143] Silva, I. S.; Gomide, J.; Veloso, A.; Meira, Jr., W. & Ferreira, R. (2011). Effective sentiment stream analysis with self-augmenting training and demand-driven projection. Em *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, New York, NY, USA. ACM.

[144] Smith, M.; Rainie, L.; Shneiderman, B. & Himelboim, I. (2014). Mapping twitter topic networks: From polarized crowds to community clusters. Pew Research Center. Last Accessed On 07/09/2015.

[145] Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y. & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Em *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631--1642, Stroudsburg, PA. Association for Computational Linguistics.

[146] Soucy, P. & Mineau, G. W. (2005). Beyond tfidf weighting for text categorization in the vector space model. Em *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, San Francisco, CA, USA.

[147] Spearman, C. (1987). The proof and measurement of association between two things. By C. Spearman, 1904. *The American journal of psychology*, 100(3-4):441--471.

[148] Stoyanov, V. & Cardie, C. (2008). Annotating topics of opinions. Em *LREC*. European Language Resources Association.

[149] Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, 10(2):175--195.

[150] Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M. & Li, P. (2011). User-level sentiment analysis incorporating social networks. Em *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pp. 1397--1405, New York, NY, USA. ACM.

[151] Tang, J.; Chang, S.; Aggarwal, C. & Liu, H. (2015). Negative link prediction in social media. Em *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pp. 87--96, New York, NY, USA. ACM.

[152] Tauer, J. (2009). Monday morning quarterbacking: The case of the hindsight bias. Psychology Today.http://www.psychologytoday.com/blog/goal-posts/200911/monday-morning

[153] Tong, H.; Faloutsos, C. & Pan, J.-Y. (2008). Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.*, 14(3):327--346.

[154] Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Em *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*.

[155] Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. Em *Proceedings of the 40th Annual Meeting on Assoc. for Computational Linguistics*, ACL '02, Morristown, NJ, USA. Assoc. for Computational Linguistics.

[156] van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, Utrecht. http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm.

[157] Volkova, S.; Wilson, T. & Yarowsky, D. (2013). Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. Em *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 505--510, Sofia, Bulgaria. Association for Computational Linguistics.

[158] Vuong, B.-Q.; Lim, E.-P.; Sun, A.; Le, M.-T.; Lauw, H. W. & Chang, K. (2008). On ranking controversies in wikipedia: Models and evaluation. Em *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pp. 171--182, New York, NY, USA. ACM.

[159] Vydiswaran, V. G. V.; Zhai, C.; Roth, D. & Pirolli, P. (2012). Biastrust: teaching biased users about controversial topics. Em wen Chen, X.; Lebanon, G.; Wang, H. & Zaki, M. J., editores, *CIKM*, pp. 1905–1909. ACM.

[160] Wallace, B. (2013). Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, pp. 1–17.

[161] Walton, D. (1991). Bias, critical doubt, and fallacies. Number 28, pp. 1–22. Argumentation and Advocacy.

[162] Watts, D. J. (2011). *Everything Is Obvious: *Once You Know the Answer*. Crown Business.

[163] Waugh, A. S.; Pei, L.; Fowler, J. H.; Mucha, P. J. & Porter, M. A. (2009). Party polarization in congress: A social networks approach.

[164] Weber, I.; Garimella, V. R. K. & Batayneh, A. (2013). Secular vs. islamist polarization in egypt on twitter.

[165] Whang, J. J.; Gleich, D. F. & Dhillon, I. S. (2013). Overlapping community detection using seed set expansion. Em *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, pp. 2099--2108, New York, NY, USA. ACM.

[166] Whannel, G. (1998). Reading the sports media audience. *MediaSport*, pp. 221--232.

[167] Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69--101.

[168] Wind, D. K. (2014). Concepts in predictive machine learning: A conceptual framework for approaching predictive modelling problems and case studies of competitions on kaggle. Master's thesis, Technical University of Denmark, Copenhagen, Denmark.

[169] Wong, F. M. F.; Tan, C. W.; Sen, S. & Chiang, M. (2013). Quantifying political leaning from tweets and retweets. Em *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*.

[170] Yang, B.; Cheung, W. & Liu, J. (2007). Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1333–1348.

[171] Yang, B.; Zhao, X. & Liu, X. (2015). Bayesian approach to modeling and detecting communities in signed network. Em *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 1952--1958.

[172] Yang, J. & Leskovec, J. (2012). Structure and overlaps of communities in networks. *CoRR*, abs/1205.6228.

[173] Yang, S.-H.; Smola, A. J.; Long, B.; Zha, H. & Chang, Y. (2012). Friend or frenemy?: Predicting signed ties in social networks. Em *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pp. 555--564, New York, NY, USA. ACM.

[174] Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Em *Proc. of the 14th Int'l Conference on Machine Learning (ICML)*.

[175] Yardi, S. & Body, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology and Society*, 5(30).

[176] Ye, J.; Cheng, H.; Zhu, Z. & Chen, M. (2013). Predicting positive and negative links in signed social networks by transfer learning. Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 1477--1488, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[177] Yin, P.; Luo, P.; Lee, W.-C. & Wang, M. (2013). Silence is also evidence: Interpreting dwell time for recommendation from psychological perspective. Em *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 989--997, New York, NY, USA. ACM.

[178] Yus, F. (2011). *Cyberpragmatics: Internet-Mediated Communication in Context*. Pragmatics & beyond new series. John Benjamins Publishing Company.

[179] Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473.

[180] Zhang, P.; Zhu, X. & Shi, Y. (2008a). Categorizing and mining concept drifting data streams. Em *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 812--820, New York, NY, USA. ACM.

[181] Zhang, Y.; Friend, A. J.; Traud, A. L.; Porter, M. A.; Fowler, J. H. & Mucha, P. J. (2008b). Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications*, 387(7):1705--1712.

[182] Zhu, X. & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation.

[183] Zhu, X. & Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.

[184] Zliobaite, I.; Bifet, A.; Holmes, G. & Pfahringer, B. (2011). Moa concept drift active learning strategies for streaming data. *Journal of Machine Learning Research*, 17:48–55.