

**DETECÇÃO DE INFLUÊNCIA NO TWITTER  
BASEADA EM SENTIMENTO**



CAROLINA ANDRADE SILVA BIGONHA

**DETECÇÃO DE INFLUÊNCIA NO TWITTER  
BASEADA EM SENTIMENTO**

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

ORIENTADOR: MIRELLA M. MORO  
COORIENTADOR: MARCOS ANDRÉ GONÇALVES

Belo Horizonte  
12 de março de 2012



CAROLINA ANDRADE SILVA BIGONHA

**SENTIMENT-BASED INFLUENCE DETECTION  
ON TWITTER**

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais - Departamento de Ciência da Computação in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: MIRELLA M. MORO  
CO-ADVISOR: MARCOS ANDRÉ GONÇALVES

Belo Horizonte

March 12, 2012

© 2012, Carolina Andrade Silva Bigonha.  
Todos os direitos reservados.

Bigonha, Carolina Andrade Silva.  
B594d Detecção de influência no Twitter baseada em  
sentimento / Carolina Andrade Silva Bigonha. — Belo  
Horizonte, 2012  
xxiii, 97 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de  
Minas Gerais - Departamento de Ciência da  
Computação

Orientador: Mirella M. Moro

Coorientador: Marcos André Gonçalves

1. Computação - Teses. 2. Redes sociais on-line -  
Teses. I. Orientador II. Coorientador III. Título.

519.6\*04(043)



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

## FOLHA DE APROVAÇÃO

Detecção de influência no twitter baseada em sentimento

**CAROLINA ANDRADE SILVA BIGONHA**

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MIRELLA MOURA MORO - Orientadora  
Departamento de Ciência da Computação - UFMG

PROF. MARCOS ANDRÉ GONÇALVES - Co-orientador  
Departamento de Ciência da Computação - UFMG

PROF. DENILSON BARBOSA  
Departamento de Ciência da Computação - UA

PROF. GISELE LOBO PAPPA  
Departamento de Ciência da Computação - UFMG

PROF. JUSSARA MARQUES DE ALMEIDA  
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 12 de março de 2012.





# Acknowledgments

I am deeply grateful to my supervisors, Mirella and Marcos. Their knowledge and guidance were essential to the development of this work. Additionally, I would like to show my gratitude to Professor Denilson Barbosa, for receiving me at the University of Alberta and for the insightful meetings. I also wish to thank my partners (and friends) at Zahpee, for whom I have great regard. Finally, my warm thanks are due to my parents Mariza and Roberto, to my sister Patrícia, to Thiago and to my dear friends for the love and support.

This work is partially supported by the project INCT-Web (MCT/CNPq grant 57.3871/2008-6) and by a FAPEMIG scholarship. This financial support is gratefully acknowledged.



*“We perceive and are affected by changes too subtle to be described.”*  
(Henry David Thoreau)



# Resumo

O conteúdo gerado por usuários e disponível em comunidades *online* é fácil de criar e consumir. Ultimamente, esse tipo de conteúdo se tornou estrategicamente importante para empresas interessadas em obter *feedback* da população em relação a produtos, propagandas, etc. Uma das comunidades *online* mais importantes atualmente é o Twitter: estatísticas recentes reportam 65 milhões de novos tweets por dia. No entanto, além de processar todo o conteúdo disponível nessas comunidades ser uma tarefa custosa, uma parte considerável dos dados não é útil para análise estratégica. Neste contexto, com o objetivo de filtrar os dados a serem analisados, propõe-se um novo método para ordenar os usuários mais influentes no Twitter de acordo com um determinado tópico. Esta nova abordagem é baseada na combinação de três fatores relacionados a cada usuário: seu relacionamento com seus vizinhos, a polaridade das suas opiniões e as características textuais dos seus tweets. A avaliação experimental deste trabalho mostra que a abordagem proposta pode, com sucesso, identificar alguns dos usuários mais influentes para três bases diferentes. Especificamente, avalia-se, para tais bases, o desempenho do método apresentado focando-se em interações entre usuários via tweets e em conexões explícitas; estuda-se o impacto de cada fator de um usuário no seu nível de influência; compara-se o desempenho do método apresentado com diversos *baselines*; e discute-se o impacto da análise automática do sentimento dos tweets na detecção de evangelistas e difamadores.



# Abstract

The user generated content available in online communities is easy to create and consume. Lately, it also became strategically important to companies interested in obtaining population feedback on products, merchandising, etc. One of the most important online communities is Twitter: recent statistics report 65 million new tweets each day. However, processing this amount of data is very costly and a big portion of the content is simply not useful for strategic analysis. Thus, in order to filter the data to be analyzed, we propose a new method for ranking the most influential users in Twitter. This new approach is based on a combination of the users' position in networks that emerge from Twitter relations, the polarity of their opinions and the textual characteristics of their tweets. Our experimental evaluation shows that our approach can successfully identify some of the most influential users on three different datasets. Specifically, we evaluate the performance of the presented method focusing on the interactions between users and the explicit connections between them; we study the impact of each perspective of users' behavior on their level of influence; we compare the performance of the presented method with distinct baselines; and, finally, we discuss the impact of automatic analysis of tweets' sentiment on finding evangelists and detractors.





# List of Figures

3.1	SaID overview. . . . .	17
3.2	Example of each interaction via tweets. . . . .	21
3.3	Example of positive tweets about Paypal. . . . .	25
3.4	Example of negative tweets about Paypal. . . . .	25
3.5	Example of neutral tweets about Paypal. . . . .	26
4.1	CCDF probabilities of following and followers for each dataset. . . . .	33
4.2	The relation between the number of followers and following of a user. . . . .	34
4.3	Quantity of tweets posted by users. . . . .	36
4.4	User data and analysis. . . . .	38
4.5	Evaluation pool results. . . . .	40
4.6	Example of plot for $recall @ x$ , $10 \leq x \leq 150$ . . . . .	42
4.7	Values of $recall @ x$ using each baseline, for <i>soda</i> dataset. . . . .	46
4.8	Values of $recall @ x$ using each baseline, for <i>appliance</i> dataset. . . . .	47
4.9	Values of $recall @ x$ using each baseline, for <i>groceries megastore</i> dataset. . . . .	48
4.10	Graphic representation of $G_i$ and $G_c$ for the datasets. . . . .	49
4.11	Paired observations for Interaction and Connection Graph approaches for evangelists and detractors' $recall @ x$ in both datasets. The parameters $(\alpha, \beta, \gamma)$ of Formula 3.5 are calculated using a <i>leave-one-out</i> procedure. . . . .	52
4.12	Plot of $recall @ x$ , using $G_i$ , considering only polarity, relation and content in both datasets. For <i>polarity</i> the parameters of Formula 3.5 are be $\alpha = 1, \beta = \gamma = 0$ , for <i>relation</i> , $\beta = 1, \alpha = \gamma = 0$ and for <i>content</i> , $\gamma = 1, \alpha = \beta = 0$ . Baseline curves are also displayed for each case, for comparison. . . . .	55
4.13	Plot of $recall @ x$ , using $G_i$ , considering only polarity, relation and content in both datasets. For <i>polarity</i> the parameters of Formula 3.5 are be $\alpha = 1, \beta = \gamma = 0$ , for <i>relation</i> , $\beta = 1, \alpha = \gamma = 0$ and for <i>polarity</i> , $\gamma = 1, \alpha = \beta = 0$ . Baseline curves are also displayed for each case, for comparison. . . . .	56

4.14	Plot of $recall @ x$ , using $G_i$ , considering only polarity, relation and content in both datasets. For <i>polarity</i> the parameters of Formula 3.5 are be $\alpha = 1, \beta = \gamma = 0$ , for <i>relation</i> , $\beta = 1, \alpha = \gamma = 0$ and for <i>polarity</i> , $gamma = 1, \alpha = \gamma = 0$ . Baseline curves are also displayed for each case, for comparison.	57
4.17	Leave one out for influence score. . . . .	63
4.18	Comparison of SaID with the baselines for evangelists and detractors. . . .	66
5.1	The 10-fold cross validation technique and final prediction of tweets. . . . .	70
5.2	The 10-fold cross validation technique and final prediction of tweets. . . . .	71
5.3	Comparing automatic and manual approaches. . . . .	74
A.1	Term cloud for positive, negative and neutral tweets for the soda dataset. . .	84
A.2	Term cloud for positive, negative and neutral tweets for the appliance dataset. .	85
A.3	Term cloud for positive, negative and neutral tweets for the groceries dataset. .	86

# List of Tables

1.1	Twitter numbers and statistics (Extracted from Twitter Blog Penner [2011]) . . . .	2
2.1	Glossary of Twitter terms. . . . .	8
3.1	Characteristics of Influential Users. . . . .	16
3.2	Example of queries. . . . .	19
3.3	Construction of Connection and Interaction Graphs. . . . .	22
4.1	Datasets' characteristics. . . . .	32
4.2	Tweets and users per sentiment. . . . .	35
4.3	Number of influential users (evangelists and detractors) in each dataset and the fraction that they represent. . . . .	41
4.4	Notation used for $recall @ x$ and $\mathcal{F}_2 @ x$ measures $m$ calculated for the polarity $i = \{e, d\}$ for each dataset. . . . .	45
4.5	Number of observations needed for $recall$ and $\mathcal{F}_2$ random plots, with 20% of accuracy and a level of confidence of 80%. . . . .	45
4.6	Statistics for $G_i$ and $G_c$ for each dataset. . . . .	48
4.7	$\mathcal{F}_2^i$ values for the ranked lists. The arrows indicates the higher (best) ( $\blacktriangle$ ) and lower (worst) ( $\blacktriangledown$ ) values. The circle ( $\bullet$ ) indicates equal or approximated values. The parameters $\alpha$ , $\beta$ and $\gamma$ used in this experiment were determined in a <i>leave-one-out</i> procedure. . . . .	50
4.8	Difference of recall ( $G_i - G_c$ ), with 90% confidence intervals. The symbol $\blacktriangle$ highlights the cases in which $G_i$ is better, $\blacktriangledown$ highlights when $G_c$ is better and $\bullet$ shows the cases in which the difference between the approaches is not statistically significant. . . . .	51
4.9	Computing time comparison, in seconds, of betweenness and eigenvector centrality in $G_i$ and $G_c$ . The symbol $\blacktriangle$ accounts for the cases in which $G_c$ have a higher time of execution than $G_i$ and $\bullet$ for the cases in which the difference includes zero and both approaches have statistically equal time of execution. . . . .	54

4.10	Factorial design results for both evangelist (E) and detractors (D) for both datasets. . . . .	60
4.11	Result of the paired observation of SaID with each baseline. We show the mean of the differences with their standard deviation and 90% confidence intervals. . . . .	64
4.12	Mean, standard deviation, minimum and maximum values for number of neutral tweets, positive tweets an followers for the evangelists. . . . .	65
5.1	Tweet automatic sentiment classification results (with the 90% confidence interval). . . . .	72
5.2	Confusion matrix of the user polarity attribution. Each column represents the users whose polarity was calculated based on the automatic classification. Each row represents the instances whose polarity was calculated using the manual classification. . . . .	73
5.3	Values of <i>precision</i> , <i>recall</i> , $\mathcal{F}_1$ and Macro- $\mathcal{F}_1$ for profile attribution using automatic classification. . . . .	73

# Contents

<b>Acknowledgments</b>	<b>ix</b>
<b>Resumo</b>	<b>xiii</b>
<b>Abstract</b>	<b>xv</b>
<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 (Electronic) Word of Mouth . . . . .	1
1.2 Opinion Leaders, Influential Users . . . . .	3
1.3 Main Contributions . . . . .	5
1.4 Text Organization . . . . .	5
<b>2 A Closer Look at Twitter</b>	<b>7</b>
2.1 Twitter Basic Concepts . . . . .	7
2.2 Twitter’s Environment and Data . . . . .	9
2.3 User Influence on Twitter . . . . .	11
2.3.1 Related Work . . . . .	11
2.3.2 Our Work . . . . .	13
2.4 Concluding Remarks . . . . .	14
<b>3 Sentiment-Based Influence Detection</b>	<b>15</b>
3.1 What is influence and how to measure it on Twitter? . . . . .	15
3.2 SaID Overview . . . . .	17
3.3 Pre-processing . . . . .	18
3.3.1 Topic and time interval definition . . . . .	18
3.3.2 Crawling . . . . .	19

3.3.3	Tweet filtering and storage . . . . .	19
3.3.4	User Data Extraction . . . . .	19
3.4	Feature Extraction . . . . .	20
3.4.1	Relation Features . . . . .	20
3.4.2	Polarity Features . . . . .	24
3.4.3	Content Features . . . . .	27
3.5	Influence Score . . . . .	28
3.6	Concluding Remarks . . . . .	29
<b>4</b>	<b>Experiments with Manual Assessment of Tweets' Sentiment</b>	<b>31</b>
4.1	Datasets . . . . .	31
4.1.1	Dataset Characteristics . . . . .	32
4.1.2	Tweet Sentiment Classification . . . . .	34
4.1.3	Influential Users: Ground Truth . . . . .	36
4.2	Experiment Setup . . . . .	41
4.2.1	Evaluation Metrics . . . . .	41
4.2.2	Baselines . . . . .	43
4.2.3	Analysis of the Baselines . . . . .	44
4.3	Interaction $\times$ Connection Approaches . . . . .	47
4.3.1	Plotting the Graphs . . . . .	48
4.3.2	Interaction $\times$ Connection-based Influence Detection . . . . .	49
4.4	Perspective Impact and Parameter Estimation . . . . .	54
4.4.1	The Impact of Each Perspective . . . . .	54
4.4.2	Estimating the Parameters . . . . .	61
4.5	Evangelists $\times$ Detractors . . . . .	63
4.6	Concluding Remarks . . . . .	66
<b>5</b>	<b>Experiments: Towards a Fully Automatic Approach</b>	<b>69</b>
5.1	Automatic Classification of Tweets . . . . .	69
5.2	From Tweets to Users . . . . .	72
5.3	Manual $\times$ Automatic Classification . . . . .	74
5.4	Concluding Remarks . . . . .	75
<b>6</b>	<b>Conclusion</b>	<b>77</b>
6.1	Contributions . . . . .	77
6.2	Future Work . . . . .	79
	<b>Appendix A Characterizing the Content</b>	<b>81</b>

A.1 Term Cloud . . . . .	81
A.2 Datasets' Example Tweets . . . . .	82
<b>Bibliography</b>	<b>87</b>





# Chapter 1

## Introduction

### 1.1 (Electronic) Word of Mouth

Word of mouth (WOM), defined as "*oral often inadvertent publicity*" by Merriam-Webster, is commonly known as the process of transferring information from person to person. Several studies in consumer behavior [Brown and Reingen, 1987, Engel et al., 1969, Katz et al., 1955] show that WOM communication is more effective in influencing consumers' attitudes than mass media, such as television, radio and newspapers. Indeed, WOM is perceived by consumers as more reliable, credible and trustworthy compared to firm-initiated communications, as stated in Schiffman and Kanuk [1999].

As interpersonal communication environments evolved to online venues, consumers started to engage in the called **electronic** word of mouth (eWOM), offering and gathering unbiased product information on the Web [Hennig-Thurau et al., 2004]. eWOM may take place in news groups, discussion forums, opinion platforms, online social networks or other environments that allow the creation of *user-generated content*. Nielsen's Global Online Consumer Survey of 2009 [The Nilsen Company, 2009] showed that 90% of the Internet consumers worldwide trust recommendations from people they know, while 70% trust consumer opinions posted online.

Among the broad variety of user-generated content environments, such as question-answer databases, blogs, digital videos, podcasts, forums, review sites, social networks, wikis and so on, **Twitter**<sup>1</sup> stands out for its simplicity and diversity. Twitter is a micro-blogging tool that represents a real-time information network. Motivated by the question "*What's happening?*", users of Twitter post messages of up to 140 characters, called *statuses*, or more familiarly, *tweets*. A tweet may include, be-

---

<sup>1</sup>**Twitter.** <http://www.twitter.com/>

Table 1.1: Twitter numbers and statistics (Extracted from Twitter Blog Penner [2011])

<b>#tweets</b>	<b>3 years, 2 months and 1 day.</b> The time it took from the first Tweet to the billionth Tweet. <b>1 week.</b> The time it now takes for users to send a billion Tweets. <b>50 million.</b> The average number of Tweets people sent per day, one year ago. <b>140 million.</b> The average number of Tweets people sent per day, in the last month. <b>177 million.</b> Tweets sent on March 11, 2011. <b>456.</b> Tweets per second (TPS) when Michael Jackson died on June 25, 2009 (a record at that time). <b>6,939.</b> Current TPS record, set 4 seconds after midnight in Japan on New Year's Day.
<b>#accounts</b>	<b>572,000.</b> Number of new accounts created on March 12, 2011. <b>460,000.</b> Average number of new accounts per day over the last month. <b>182%.</b> Increase in number of mobile users over the past year.

sides pure text, links to websites, photos, videos and other media, as well short strings preceded by a hash symbol (`#`), called *hashtags*, usually employed to filter or promote content [Huang et al., 2010].

One of the main characteristics of Twitter is that, due to the message short size and the effortless posting / reading from anywhere, it is easy to both produce and consume content. Jansen et al. [2009], indicate this immediacy of posting (one can send a tweet at the moment of a purchase or a problem in the bank) and the simplicity of finding out what people are talking about as the main factors why Twitter plays a major role in eWOM. Moreover, O'Connor and Balasubramanian [2010] show that text streams (such as Twitter) are a potential substitute and supplement for traditional public opinion surveys.

In summary, opinions, experiences and suggestions are shared by users on Twitter in large scale. Considering the Twitter users as potential consumers / voters and the WOM generated by their discussions, micro-blogging networks have become a rich source of data in any situation in which feedback is desired. Reacting properly to the information available in Twitter has become essential for businesses [Brown et al., 2007]. By studying the data and the users, they can gather market intelligence and improve their campaigns, products or services acceptance.

Analyzing this data is not simple, though, due to the huge amount of content generated daily. For example, Table 1.1 presents some numbers reported on Twitter Blog Penner [2011] that illustrate Twitter's growth. Besides being impractical to inspect all the data generated daily (even for a specific topic), not all tweets and users are worth such an evaluation. Under these circumstances, it is crucial to find the opinion leaders, or *influential users*, who drive eWOM conversations on Twitter. By targeting these key users, marketers can benefit from a social multiplier effect on their marketing efforts [Van den Bulte and Joshi, 2007] and leverage lower (and strategic) investments [Slywotzky and Shapiro, 1993].

## 1.2 Opinion Leaders, Influential Users

Katz et al. [1955] defined as *opinion leaders* “the individuals who were likely to influence other persons in their immediate environment”. Although some (e.g., Watts and Dodds [2007]) may question the existence of opinion leaders (or “influentials” as they are also called [Merton, 1968]), their presence and importance are widely discussed on marketing and business environments [Barabasi, 2002, Berry and Keller, 2003, Chan and Misra, 1990, Evangelopoulos and Visinescu, 2012, Gladwell, 2002, Slywotzky and Shapiro, 1993, Van den Bulte and Joshi, 2007]. According to Chan and Misra [1990], the propagation of information through word of mouth communication makes opinion leaders prominent among their group. Their leadership, which may be an indication of innovativeness, comes from their persuasion as early adopters towards later adopters to try a new product or service. This happens, as stated in [Katz et al., 1955, Lazarsfeld et al., 1948], because in a variety of decision-making scenarios, individuals may be influenced more by exposure to each other than to the media.

Assuming the existence of such influential users, we explore what we call *sentiment-based influence given a topic*. The focus on topics is because people are often interested in monitoring one particular topic or context, e.g., a product, a personality, an event [Savage, 2011]. And it is sentiment-based motivated by insights that can be extracted from polarized content [Arndt, 1967, Chevalier and Mayzlin, 2006, Diakopoulos and Shamma, 2010, O’Connor and Balasubramanyan, 2010]. We perform an analysis focusing on positively and, especially, negatively biased users: as shown in Lee et al. [2008], Mizerski [1982], negative online reviews have a more powerful impact on product attitude than the positive ones. The intuition is that negative posts are more likely to induce consumers to change their mind about a product (and choose another one) than positive ones. Moreover, on a crisis manager perspective, identifying negatively biased users may simplify the marketing analysis for branding strategy and brand-customer interaction.

We define the characteristics expected on influential users and propose an approach for finding them on Twitter. Specifically, we formalize influential users as the well connected ones who produce content with potential for changing people’s opinions. In other words, influential users are those:

- (i) **who are convincing positively or negatively;**
- (ii) **who act like bridges in interactions among a subject;**
- (iii) **whose actions imply in other’s actions;**

(iv) **whose content satisfies a minimum expected quality.**

Once we have defined the influential user general profile, the question is then how to identify those users among all. Hence, the method presented in this work for identifying influential users is based on three main perspectives that summarize the behavioral profile of a user on Twitter: polarity, relation and content. First, the polarity features, calculated based on the classification of tweets, consider the user overall contribution to the topic discussion. In other words, the content must be either positive, neutral or negative such that we can classify the user as an evangelist or a detractor. Second, for the relation features, we capture network properties that represent user's interactions around a topic. Based on such network properties, we apply centrality metrics to rank the notoriety of users according to their position in the network. Influential users have to be well connected to other users, and play a central role in the graph in which they are embedded. Finally, we study content features of the user. We hypothesize that if users are to influence other people, their tweets are expected to have a minimum quality. As shown by Brown et al. [2007], consumers seem to evaluate the credibility of online WOM information in relation to the individual contributor of that information. Content features correspond to the analysis of the readability of the tweet content, ranking higher posts (and, consequently, their authors) that are well written and understandable according to readability metrics.

With that information at hand, the next step is to rank the users according to their level of influence. In order to do so, we have also defined an Influence Score, that combines all user features into one single factor.

For testing our techniques, we built three datasets for specific topics (two product brands and a groceries megastore chain). Each tweet was manually classified as *positive* / *negative* / *neutral*. Also, each user was categorized as *evangelist* / *detractor* / *not influential*. We used this categorization as ground truth experiments.

Our experimental results demonstrate that we can successfully identify some of the most influential users concerning a subject using our techniques. We also show that interactions between users are a better evidence to determine user influence than explicit connections and that the automatic classification of the tweets does not impact much the results for influence detection. The experiments were performed in diverse topic-specific scenarios, demonstrating the applicability of the method to diverse subjects. Moreover, we show that the topic-specific datasets employed have similar characteristics when compared to some more general Twitter collections used in previous work, such as Huberman et al. [2008] and Krishnamurthy et al. [2008], meaning that most of our results are potentially generalizable.

## 1.3 Main Contributions

The main contributions of this work are summarized as follows.

- (a) A new and clear definition of what an influential user is;
- (b) A method, called SaID (Sentiment-based Influence Detection on Twitter) for detecting influential users based on the polarity of their tweets;
- (c) Fully analyzed datasets (users and tweets) that can be used as benchmark for future work;
- (d) Detailed comparison of the effect of interactions via tweets and follower/following on influence detection;
- (e) Detailed evaluation of the contribution of polarity, relations features and content quality on influence detection;
- (f) Analysis of the impact of an automatic tweet sentiment analysis on influence detection.

Contributions (a) to (e) are also published in the following papers:

- **Detecting Evangelists and Detractors on Twitter.** C. Bigonha, T. N. C. Cardoso, M. M. Moro, V. Almeida, and M. A. Gonçalves. *In Brazilian Symposium on Multimedia and the Web (WebMedia)*, 2010.
- **Sentiment-based Influence Detection on Twitter.** Carolina Bigonha, Thiago N. C. Cardoso, Mirella M. Moro, Marcos A. Gonçalves and Virgílio A. F. Almeida, *Journal of the Brazilian Computer Society*, 2011. DOI:10.1007/s13173-011-0051-5.

## 1.4 Text Organization

This work is organized as follows. Firstly, Chapter 2 overviews Twitter characteristics as well as the related work. Chapter 3 presents SaID (Sentiment-based Influence Detection), our method for influential users identification. Chapter 4 presents our experimental evaluation and discusses the main results. Chapter 5 compares SaID results using manual and automatic classification of the tweets. The purpose of this last experiment is to evaluate the impact of the automatic classification on our influence detection method. Finally, Chapter 6 concludes this dissertation, reviewing our main contributions and discussions.



# Chapter 2

## A Closer Look at Twitter

In this Chapter, Twitter's main characteristics and features are visited in order to facilitate the understanding of this dissertation. We describe basic concepts of Twitter (along with a glossary table) and several studies that aim to characterize the users and data of Twitter. Then, an overview of recent work concerning influence on Twitter is presented.

### 2.1 Twitter Basic Concepts

Twitter is a *micro-blogging service* and a *real-time information network*, in which content is shared between users through short length text-based posts called *tweets*. Twitter was created in early 2006 [Arrington, 2006, Penner, 2011], and its popularity has increased quickly [Weil, 2010]. The number of Twitter users is estimated in about 200 million [Shiels, 2011]; only on March 12, 2011, 572 thousand new accounts were created [Penner, 2011]. As this service evolved, both the research and business communities became more interested in it and plenty analysis of this environment's users and data took place, as discussed next.

Twitter asks its users the question "*What's happening?*", allowing them to answer in quick and frequent *Twitter status updates*, or tweets. A tweet is a text-only message with at most 140 characters. It can be sent through Twitter's website, SMS, instant messaging, email, mobile devices and desktop applications.

In addition to posting tweets, users of Twitter can follow each other. Following others on Twitter means subscribing to their tweets as a follower. To follow a user is a unilateral action. Users can follow other users that do not follow them back.

Lists of tweets arranged in real-time order are called *timeline*. All users have a home timeline in which their tweets and new tweets from the people they follow appear

in real time, as the content is created. Newest messages are at the top. Besides the timeline of tweets, the users' profile pages display their respective number of posted tweets, number of followers, number of people they are following, a picture, a short bibliography, the user's provided real name, and so on.

Table 2.1: Glossary of Twitter terms.

TERM	DEFINITION
<b>User</b>	Each user in Twitter has a unique username, a profile page, and a set of tweets.
<b>Tweet</b>	Messages with 140 characters or less posted to Twitter. It appears on the sender's profile page and in the home timeline of anyone who is following the sender.
<b>Timestamp</b>	The time and day the tweet was posted. It appears within each tweet.
<b>Timeline</b>	Any ordered real-time list of tweets. It appears on profile pages (the list of tweets posted by the respective user), on users' home page (the list of tweets they have posted along with the tweets the people they follow posted), as the results of a search, and so on.
<b>Profile page</b>	Users' profile page contains their personal data, such as picture, real name and description. Also it contains their timeline and some quantitative data, such as follower count, following count, number of tweets posted, and so on.
<b>To Follow</b>	To subscribe to another user's tweets.
<b>Follower</b>	A follower of a user is one that is subscribed to her tweets.
<b>Friend (Following)</b>	A friend of a user is one who is followed by her.
<b>Protected Account</b>	The tweets from protected account users are only seen by approved followers and they do not appear on searches.
<b>@Mention</b>	Tweets containing other user's username, preceded by the "@" (at mark) symbol. It appears on the sender's Profile page and in the recipient's @Mentions tab, in Twitter home page. If the recipient is following the sender, it will also appear in the recipient's timeline. @Mention tweets usually have the format '<content> @username <content>'. </td></tr> <tr&gt; "direct="" "hottest="" "retweet="" "rt"="" "sent"="" #="" '&lt;="" (#)&lt;="" (api).="" (identified="" <="" <td&gt;&lt;b&gt;@reply&lt;="" <td&gt;&lt;b&gt;direct="" <td&gt;&lt;b&gt;geotagging&lt;="" <td&gt;&lt;b&gt;hashtag="" <td&gt;&lt;b&gt;retweet&lt;="" <td&gt;&lt;b&gt;trending="" <td&gt;&lt;b&gt;twitter="" <td&gt;&lt;p&gt;a="" <td&gt;&lt;p&gt;private="" <td&gt;&lt;p&gt;some="" <td&gt;&lt;p&gt;the="" <td&gt;&lt;p&gt;twitter="" <td&gt;&lt;p&gt;words="" <tr&gt;="" @mentions="" @reply="" a="" act="" again="" algorithm).="" already="" also="" an="" analysis.&lt;="" and="" another="" anywhere="" api&lt;="" appear="" appears="" application="" applications="" are="" attached="" b&gt;&lt;="" beginning="" begins="" build="" by="" called="" can="" considered="" content="" data="" developers="" directly="" discussion="" div="" emerging="" enable="" enables="" exposed="" filtering="" folder="" folder.="" follow.="" followers="" following="" follows="" for="" format="" generated="" geolocation="" goal="" happens="" hash-tags="" have="" home="" icon"="" identified="" if="" immediately="" in="" interface="" is="" it="" keywords="" landing="" letters="" message="" messages"="" messages&lt;="" not="" occur="" of="" on="" one="" only="" or="" p&gt;&lt;="" page="" page.="" page.&lt;="" people="" permit="" popular="" post="" posted="" posted.&lt;="" preceded="" present="" profile="" programming="" provides="" public.&lt;="" published.="" querying="" recipient="" recipient's="" reply="" retweet="" search&lt;="" search,="" sender's="" sender,="" sender.="" sent="" share="" specifying="" symbol,="" tab,="" table&gt;="" tag="" tbody&gt;="" td&gt;="" td&gt;&lt;="" that="" the="" their="" they="" timeline.="" timelines.&lt;="" to="" tool,="" topic="" topic&lt;="" topics="" tr&gt;="" trending="" trends="" tweet="" tweet.="" tweet.&lt;="" tweets="" tweets.="" tweets.&lt;="" twitter="" twitter".="" twitter's="" user="" user's="" username="" users="" usually="" via="" was="" when="" where="" will="" with=""> <div data-bbox="181 1928 1369 2018" data-label="Text"> <p>When users sign up for Twitter, they have the option of keeping their tweets public (the default option) or protected. Accounts with public tweets have their profile pages</p> </div></tr&gt;>



visible to everyone. In addition, their tweets can be searched using Twitter Search<sup>1</sup> or retrieved using Twitter API<sup>2</sup>. On the other hand, for accounts with protected tweets, a manual approval is required for each person who may want to see the tweets. Tweets posted on these accounts are visible only for the approved ones.

Users may interact with each other via tweets in different forms, such as *@Replies*, *@Mentions* and *Retweets*. These forms of interaction are considered variations of a normal tweet and all of them contain references to other users (their usernames preceded by an @ – *at* mark) in the post. These references occur in different ways and with different purposes: *@Replies* consists on sending tweets as reply to one user, *@Mentions* consists on mentioning a user in the middle of one tweet, and *Retweets* consists on sending a tweet already posted by another user.

Table 2.1 serves as a glossary for future reference and summarizes some of these definitions and other Twitter terms that may be important for understanding this work. More information concerning Twitter vocabulary and definitions can be found at Twitter Help Center<sup>3</sup>.

## 2.2 Twitter's Environment and Data

Twitter drew the attention of several researchers in the last years. Among the wide range of studies about the micro-blogging tool, there are both characterization- and application-focused studies about Twitter's environment, users and data that are worth mentioning. This section summarizes the main contributions of some of these studies.

Although many people may consider Twitter as a social network service, it actually is not. Twitter facilitates social networking, but it does not necessarily act as a social networking website. A *social network* is a social structure made of nodes (actors, individuals, organizations) and ties (relationships) among them [Barnes, 1954, Bornholdt and Schuster, 2003, Wasserman and Faust, 1994]. In actual online social networks, such as Facebook and Flickr<sup>4</sup>, the relation between the users (friendship) is reciprocal. When a user adds another one as a friend, both sides share this connection. On the other hand, the possibilities of connections provided by Twitter are not necessarily mutual: following, mentioning and retweeting are one-way actions. For example, one may opt to receive the updates from another user without requiring mutual following from that user.

---

<sup>1</sup>**Twitter Search:** <http://search.twitter.com>

<sup>2</sup>**Twitter API:** <http://dev.twitter.com>

<sup>3</sup>**Twitter Help Center:** <https://support.twitter.com/groups/31-twitter-basics>

<sup>4</sup>**Facebook and Flickr:** [http://www.\[facebook, flickr\].com](http://www.[facebook, flickr].com)

To study this property of Twitter, Kwak et al. [2010] analyze over 41 million user profiles and 1.47 billion follower/following relationships to conclude that only 22% of the connections are reciprocal, whereas the majority (78%) are one-way relationships. These facts highlight Twitter's power as a content distribution platform. Users follow others seeking not only to maintain in touch with "real life" connections, but, more importantly, to get access to information and links of interest.

Still on the relationships between users, Huberman et al. [2008] define two types of public posts: direct and indirect ones. Direct posts are destined to one specific person (using an '@' - *at* mark - in front of the user's name), whereas indirect posts do not include mentions to any other user: they are destined to everyone. Using this definition, Huberman *et al.* introduce the concept of a "user's friend" as a person to whom the user has written at least two public direct posts. Their work shows that the explicit relations between users (following relation) do not correspond to the real connections between them. The number of people with whom the users interact is way lower than the number of their connections: not every link between two people implies in a real interaction between them.

The content produced by users on Twitter is highly heterogeneous and rich. Much of what is discussed in Twitter is inspired by the news: according to Kwak et al. [2010], 85% of Twitter posts are news-related. In this context, several studies explore Twitter's power as a real-time news source. For example, Cataldi et al. [2010], Sankaranarayanan et al. [2009], and Phelan et al. [2009] try to retrieve real-time breaking news (or emergent topics) from Twitter users' posts; whereas Chen et al. [2010] propose ways of filtering Twitter stream down to items that are indeed of interest of the user. Castillo et al. [2011] assess information credibility (in the sense of believability) of news spread in Twitter. They automatically determine which topics are newsworthy and, more specifically, what their level of credibility is.

In summary, Twitter users can act as either **providers of news** (for example during the 2009 post-election protests in Iran [Zhou et al., 2010]) or as **opinion sources** about existing topics. In the former scenario, Sakaki et al. [2010] propose an algorithm for monitoring tweets for target event detection (such as an earthquake). Moreover, users' posts usually provide good insights about the impact of news events [Mathioudakis et al., 2010, Tsagkias et al., 2011] and also relevant information related to politics [Diakopoulos and Shamma, 2010, Golbeck and Hansen, 2011, Wigand, 2010] and for gathering market intelligence [Brown et al., 2007, Jansen et al., 2009, Kwon and Sung, 2011].

As any user-generated content environment, Twitter is also threatened by content polluters, malware disseminators and spammers. Lee et al. [2010b], Chu et al. [2010]

and Grier et al. [2010] tried to identify this type of behavior. Grier *et al.* found that Twitter has a higher incidence of users visiting spam pages than email, which may happen due to features unique to Twitter exploited by spammers. For example, mentions are used by spammers to personalize messages in an attempt to increase the likelihood of victim to follow a spam link. Retweets and trending topic hashtags are also exploited by spammers with similar objectives.

Finally, the constraints in Twitter posts (the length restriction, the variety of social relation types, the complex linguistic style [Danescu-Niculescu-Mizil et al., 2011]) are actual obstacles in terms of content analysis. Therefore, sentiment classification [Guerra et al., 2011, Jiang et al., 2011, Speriosu et al., 2011, Thelwall et al., 2010] as well as text classification for information filtering [Sriram et al., 2010] specific for this environment are important research topics.

## 2.3 User Influence on Twitter

The target content for studying influence is user-generated, so the characterization of the authors of the tweets in terms of their general behavior is crucial for a better understanding of the data. Specifically, as aforementioned, the identification of influential users or opinion leaders is important for marketers, businesses or other people interested in general feedback [Barabasi, 2002, Berry and Keller, 2003, Chan and Misra, 1990, Gladwell, 2002, Slywotzky and Shapiro, 1993, Van den Bulte and Joshi, 2007].

In this section, we list the main characteristics of each relevant work studying user influence on Twitter and discuss our main contributions when compared to them.

### 2.3.1 Related Work

The report presented in [Leavitt et al., 2009] highlights interactions (replies, retweets, mentions and attributions) as markers of influence, rather than solely the number of followers. The authors select a few famous users belonging to the categories “celebrity”, “news outlet” and “social media analyst” and compare several influence indicators, such as average content spread per tweet, for each user.

A method for topic-sensitive influential users detection is defined in [Weng et al., 2010]. Considering a *Pagerank* [Brin and Page, 1998] alike metric, it calculates users’ influence based on how many people have received their tweets. As for evaluating the

results, they compare three different algorithms (number of followers, pagerank and topic-sensitive pagerank) studying the correlation between the rank lists generated.

In Cha et al. [2010], influence is divided into three types: the in-degree influence (the number of followers that a user has), the re-tweet influence (the number of re-tweets containing ones name), and mention influence (the number of times a user is mentioned). The authors study the dynamics of influence across topics and time, analyzing whether users can hold significant influence over a variety of topics, and examining the rise and fall of influential users over time.

Based on the concept that influence is measured by the replication of already performed actions, Goyal et al. [2010] propose a technique for constructing influence probability graphs from social networks (friendship graph) and action logs. From these two sources of data, the authors build a propagation graph (in which nodes are the users who perform the actions and edges represent the direction of the propagation), apply models of influence (static, discrete and continuous time) and finally construct the graph of influence probabilities. Both Goyal et al. [2010] and Lee et al. [2010a] emphasize the temporal aspect of influence detection, which is indicated as future work of this dissertation.

In Bakshy et al. [2011], the authors measure influence based on the user's ability to spread brand new content. Given a propagation path traced from the user that created the content (URL) to the last user that received it, they identify the users who are nearer to the origin as the most influent. The attributes considered for the calculation of influence are: the number of followers, number of followings, number of tweets posted and date the user joined Twitter. The authors also analyzed the content of the links posted, observing the average cascade size for different interest ratings, types and categories of posts.

Despite focusing mainly on the topological characteristics of Twitter and its power as an information sharing environment, Kwak et al. [2010] compare three methods for ranking users: the first strategy ranks users by the number of followers, the second applies PageRank to a network of followings and followers, and the third one ranks users according to their number of re-tweets. As conclusion, the authors find the same gap between the number of followers and the popularity of one's tweets indicated before.

Liu et al. [2010] define heterogeneous network as a graph in which different types of nodes are connected through directed or undirected edges. In most of the online social networks, the nodes are users or documents and the links represent friendships between users, authoring relationships between documents, and so on. The authors' goal is to perform a topic-level influence analysis and user behavior prediction in these networks. Liu et al. are the first to consider indirect influence (meaning friend-of-a-

friend influence) and topic-specific influence. They conduct experiments on Twitter, Digg and Cora (a citation network), aiming to evaluate the influence strength prediction, the user behavior prediction and the topic-level influence model.

Pal and Counts [2011] address the problem of finding topical authorities in microblogs. They list some metrics of potential authorities and group them in higher level indicators, which consider the authors' level of involvement in the topic, the originality of their tweets, their conversational level, the impact of the content the author generates, their information diffusion and so on. Pal et al. use a probabilistic clustering over the feature set and within-cluster ranking procedure to generate a final list of important authors for one topic. They explore three different topics (extracted, using keyword matching, from all tweets posted in a 5-day interval), comparing their method with three baselines (one that uses graph properties, other that uses textual properties and a third that randomly selects non-important authors). They manually evaluate the ranks generated by each aforementioned options. One interesting metric used in that paper is the *self-similarity score*, which reflects how many users borrow words from their own previous posts (concerning or not the topic).

Finally, academic researchers are not the only ones to study user influence. Klout<sup>5</sup> is a tool that measures influence online. It provides a way to measure influence on Twitter using a score also called Klout, with a range is from 0 to 100. Light users score below 20, regular users around 30, and celebrities start around 75. Klout defines influence based in three concepts: True Reach (size of the audience of a user – true followers count, total of retweets, mentions and lists in which the user is included); Amplification Probability (chance that the published content will spread – interaction between users and their followers, ratio between the number of retweets and the number of followers); and Network Influence (analysis of the audience influence considering the interaction between the followers). Even though some may question the meaning of Klout Score [Braunstein, 2011] and there is no clear explanation of how the score is calculated, it is broadly known and used [Ishida, 2011].

### 2.3.2 Our Work

Our contributions in this work stand out from previous work in key aspects. First, our proposed method, SaID, considers more complete metrics for measuring the repercussion of users' actions: we evaluate features of users within an interaction network that captures all the conversations about a topic. Second, we are the first to apply a tweet content quality analysis: our hypothesis is that users who want to influence others tend

---

<sup>5</sup>**Klout.** <http://klout.com/>

to produce better written and more understandable tweets. Also, we evaluate the commitment of the users with the topic, that is, if they have a positive or negatively biased content and with what frequency they post about the subject. This allows our method to identify the potential evangelists and detractors concerning the topic. Finally, no previous work has evaluated their proposed method using a specialists' ground truth. Instead of generating various ranked lists and simply comparing them, we validate our technique based on marketing and communication specialists' point of view and on an evaluation pool.

## 2.4 Concluding Remarks

In this Chapter we addressed the characteristics of Twitter, a modern and trendy online environment, that can be seen as a large information network. In this environment, diverse types of users post short length messages (called tweets) concerning various matters, that go from personal comments to breaking news.

It did not take long for businesses to realize that Twitter is a valuable tool for connecting to customers in a real-time basis. By engaging on this environment, entrepreneurs are able to gather market intelligence, feedback, build relationships with customers and partners as well as react properly to negative word-of-mouth. However, given the vast amount of tweets per day, even for a specific subject, sometimes it is impractical (and not strategical at all) to analyze the whole data. In order to save time (and resources) it is critical for the businesses to find opinion leaders, or influential users, who drive the conversations about the topic.

We covered, in this Chapter, the effort of several studies on retrieving influential users on Twitter. Also, we briefly introduced the advantages of our method, called SaID (Sentiment-based Influence Detection) on Twitter, which is further described in the next Chapter.

# Chapter 3

## Sentiment-Based Influence Detection

In this chapter, we explicitly define what are the characteristics of the users that we consider influential and how these features may be measured on Twitter (Section 3.1). Furthermore, we present SaID, our Sentiment-based Influence Detection method, which is based on (1) the definition of a topic of interest, (2) a crawling of the correspondent tweets, (3) the identification of the profiles, (4) the sentiment classification, interaction parsing and content analysis of the gathered data and (5) the extraction of features for associating an influence score to each user. Specifically, we discuss SaID’s three phases: pre-processing (Section 3.3); feature extraction (Section 3.4) and calculation of the influence score (Section 3.5).

### 3.1 What is influence and how to measure it on Twitter?

Given Twitter power as an intelligence source for branding strategy [Jansen et al., 2009] and the importance of uncovering the key influential users [Goyal et al., 2010, Pal and Counts, 2011, Weng et al., 2010], our first challenge was to define the characteristics of an influential user on Twitter.

The term **influence** comes from Latin *influens*, present participle of *influere*: to flow in. It means, according to Merriam Webster, “*the act or power of producing an effect without apparent exertion of force or direct exercise of command*” or “*the power or capacity of causing an effect in indirect or intangible ways*”.

In word of mouth environments, this **effect** flows from opinion leaders [Katz

Table 3.1: Characteristics of Influential Users.

Characteristic	How we measure it
<b>Their actions imply in other persons' actions.</b>	
This point is directly derived from the basic definition of influence (by Merriam Webster) and opinion leaders (by Chan and Misra [1990] and Katz et al. [1955]). A influential user is usually the one whose actions cause effect on others.	We measure this characteristic by analyzing mentions, replies and retweets of a user contextualized on the topic of matter ( <i>Relation Features, Section 3.4.1</i> ).
<b>They act like bridges on interactions about a subject.</b>	
To maintain a leadership role on a topic, the user has to be a part of the active public, of the active discussion [Chan and Misra, 1990]. Moreover, since they generate buzz around their posts [Lazarsfeld et al., 1948, Slywotzky and Shapiro, 1993], they act like bridges on interactions, that is: they are central in the discussions.	We evaluate the user position on both connection (follower-following) and interaction (mention, reply, retweet) networks. Users that are more central, play a more important role in the network. ( <i>Relation Features, Section 3.4.1</i> ).
<b>They have a positive or negative bias on their opinion.</b>	
Several insights can be extracted by analyzing a polarized content [Arndt, 1967, Chevalier and Mayzlin, 2006, Diakopoulos and Shamma, 2010, Kwon and Sung, 2011, Mizerski, 1982, O'Connor and Balasubramanyan, 2010] and, specially for marketing strategy, the management of positive and negative feedback is important. By targeting negatively influent users, marketers can react properly, leveraging less resources [Slywotzky and Shapiro, 1993].	For each Twitter user, we analyze the sentiment of each tweet and the overall user polarity. ( <i>Polarity Features, Section 3.4.2</i> ).
<b>Produces a content with a minimum quality.</b>	
By <i>minimum quality</i> , we mean well structured sentences, with the intention of presenting an idea, usually with a source (URL) [Weng et al., 2010]. Influential users are not occasionally talking about the topic, they have a purpose for posting content. Furthermore, there is indication of a positive correlation between opinion leaders and a higher level of education [Chan and Misra, 1990, Robertson and Myers, 1969].	We evaluate the content generated by each user on Twitter. We analyze the readability features, and the presence of the user on the topic (amount of tweets on the subject). ( <i>Content Features, Section 3.4.3</i> ).

et al., 1955] to regular users (or from early to later adopters) and may concern the willingness to buy or not a product, to vote or not on a candidate in elections, and so on [Arndt, 1967, Berry and Keller, 2003, Chan and Misra, 1990, Lazarsfeld et al., 1948, Van den Bulte and Joshi, 2007]. The complicated part is to effectively identify this effect through Twitter users' extractable data and actions.



Most of the studies that discuss influential users on Twitter [Bakshy et al., 2011, Cha et al., 2010, Leavitt et al., 2009, Liu et al., 2010, Pal and Counts, 2011, Romero et al., 2011, Weng et al., 2010] propose a definition of influence based on user content diffusion. Bakshy et al. [2011] and Romero et al. [2011] study influence as the ability of the user to post (as a seed) URLs which diffuse through the Twitter follower graph; Cha et al. [2010] focus on users' potential to lead others to engage in a certain act, their size of audience and their amount of content with pass-along value; Weng et al. [2010] say that the influence of users on their followers is based on the relative amount of content the followers receive from them. Finally, Pal and Counts [2011] consider the level of involvement of the user with the topic and how much the author is mentioned with regards to the topic of interest as the most important features to evaluate a influential user.

In this dissertation, we try to refine the definition of influence on Twitter, focusing on a marketing and consumer point of view [Kwon and Sung, 2011]. Table 3.1 introduces each of the key points of our definition as well as our way of measuring each corresponding characteristic on Twitter.

## 3.2 SaID Overview

In this dissertation, we present a method called SaID (**S**entiment-based **I**nfluence **D**etection on Twitter) for identifying influential users on Twitter, which relies on the aforementioned characteristics.

Figure 3.1 shows an overview of the proposed method. Our method is divided into three main phases: *pre-processing*, *feature extraction* and *influence score*. The

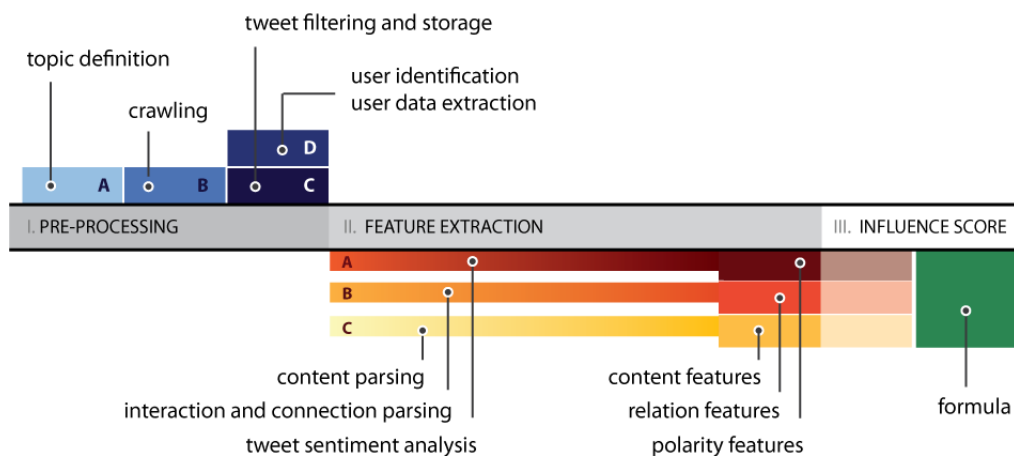


Figure 3.1: SaID overview.

first phase (*I*) consists on topic and query definitions, crawling the tweets, filtering the content, user identification and data extraction. The second phase (*II*) corresponds to the processing of the metrics: SaID parses the content, the interactions and connections between the users and analyze the sentiment of the tweets. At the end of this phase there are metrics associated with three groups of features: the ones related to **content**, **relation** and **polarity**. Finally, in the third phase (*III*), SaID combines these metrics into a single influence score.

Each of these phases are described next: Section 3.3 presents the pre-processing phase, Section 3.4 defines the features (and explains their extraction) and Section 3.5 describes how the separate features are combined into one influence score.

## 3.3 Pre-processing

The *pre-processing* phase consists of four steps. The first one is determining the topic and time interval of the desired content; the second is crawling; the third one is the tweet filtering and storage; and the fourth is the extraction of user data. This section describes each one of them.

### 3.3.1 Topic and time interval definition

In the marketing environment (considering business owners, investors and advertising agencies, for example) the interest is usually directed to a topic-restricted analysis of influence rather than a global one. For example, an important biologist is possibly not as influent as a politics-engaged user when it comes to discussing this year's election.

Under those circumstances, this work evaluates users' influence factors considering topic-related scenarios. Thus, the first step in the pre-processing phase is to determine *which topic is going to be analyzed* and *for how long*. It may be a brand, a product, a personality, an event, and so on. Based on the chosen topic, keyword-based queries are built.

Table 3.2 shows examples of the keywords-based queries used for some of the datasets evaluated in Chapter 4. It is important to notice that some knowledge about the topic of interest is required in order to produce a good query. In the case of *brastemp* (an appliance brand), the name of a line of products (*allblack*) was included as keyword, in order to improve the recall of relevant results.

A definition of the time interval is important because SaID calculates the user Influence Score based on a snapshot of the conversations for a topic. The time interval

Table 3.2: Example of queries.

Topic	Keyword-based query
<b>brastemp</b>	brastemp OR allblack OR bgourmet OR (inverse AND (geladeira OR freezer))
<b>carrefour</b>	carrefour OR carrefourbrasil OR carrefourcombr OR #carrefourto OR carrefourfail

may be as wide as desired, but has to be defined. As future work, we address SaID’s capability of adapting the user Influence Score over time.

### 3.3.2 Crawling

For collecting the data concerning the chosen topic, we use the Twitter Search API, a dedicated API for running searches against the real-time index of recent tweets. The Search API is not a complete index of all tweets, but instead an index of recent Tweets. Due to resource constraints, the results are focused in relevance and not completeness. That means that some tweets and users may be missing from search results.

We have built a crawling module, that uses Twitter Search API for collecting tweets, publicly available from the user’s timeline, which contains the defined keywords. Since Twitter Search API cannot be used to collect tweets older than a week, our crawler does real-time requests, during the desired time interval. Before storing the retrieved tweets in our database, we analyze its validity, as described in the next step.

### 3.3.3 Tweet filtering and storage

Once the content is retrieved from Twitter, we carefully eliminate occasional spams and tweets that fit into the keyword search, but in a different context. For example, on a search with the keyword “house”, there may be tweets concerning “house”, the human habitat, or “House”, the TV series. This process was conducted manually. After this filtering, the remaining tweets are stored.

### 3.3.4 User Data Extraction

Finally, the last step of the pre-processing phase is user identification and data extraction. As already mentioned, our method gathers the content generated on Twitter via tweets that match a certain query. Since our interest is on user’s characterization, we must identify the authors of the tweets and collect their information. We store the authors’ name, their profile URL and their list of followers and following users. We retrieve this information also using the Twitter API.

## 3.4 Feature Extraction

The second phase is the actual influence analysis, in which *relation*, *polarity* and *content* features of the user are extracted, as explained next.

### 3.4.1 Relation Features

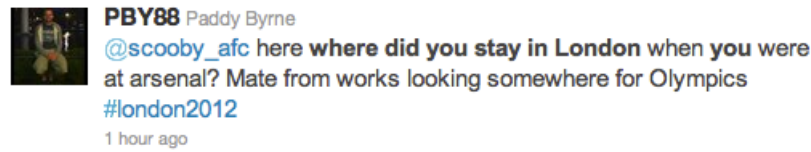
We assume that the level of influence of users is directly associated with their social relation with other users in the same topic context. Thus, the set of *relation features* tries to capture the user role among the others, in terms of follower-following connections and interactions via tweets, as described next.

#### On follower and following relations

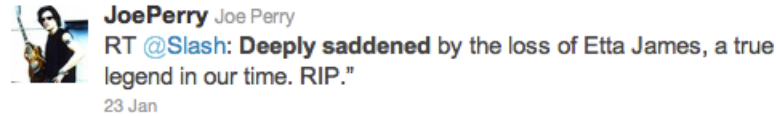
The early studies examining influence on Twitter used to confound influence and popularity, by measuring the level of influence of users by their number of followers. According to the results presented in previous work [Cha et al., 2010, Huberman et al., 2008, Weng et al., 2010], the number of followers and followees of users is not the preferred way to measure influence, because most of the users' followers do not even read or process the received posts.

However, the ratio of followers to followees, **Twitter Follower-Followee Ratio** (*tff*), may be useful as an influence indicator, because it can communicate the intended purpose of a user [Krishnamurthy et al., 2008, Leavitt et al., 2009]. According to Krishnamurthy et al. [2008] and Leavitt et al. [2009], if the ratio approaches infinity ( $\uparrow$  followers,  $\downarrow$  following), the user is likely to be a “broadcaster”, such as news media profiles, celebrities or other popular users. On the other hand, if the ratio approaches 1 (followers  $\simeq$  followees), the users have reciprocity on their connections. This describes the most common types of user. Finally, if the ratio approaches zero ( $\downarrow$  followers,  $\uparrow$  followees), the user might be categorized as a spammer or a robot, which follows way more users than is followed by (people do not usually follow back spammers/robots).

Based on such characteristics, *tff* is presented as the first relation-based metric for studying the collected data. We use this metric, combined with others, to identify influential users in our dataset, considering the users with higher *tff* as more relevant. This metric helps eliminating potential spammers (that may fit in the second and third groups) and rewards the users who are widely followed, but are selective for following others.



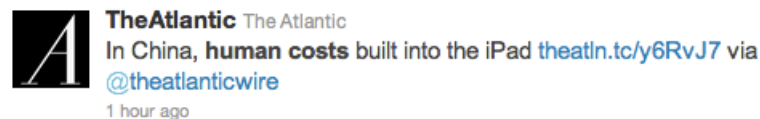
(a) Reply.



(b) Retweet.



(c) Mention.



(d) Attribution.

Figure 3.2: Example of each interaction via tweets.

### Interactions via Tweets

Previous work [Huberman et al., 2008] has shown that the number of people with whom users truly interact is inferior when compared to the number of their explicit follower-following connections: not every link between two people implies in a real interaction between them. Thus, in order to truly understand the relation between users, we analyze the *interactions that occur via tweets*.

It is very common for a user to interact with others via tweets by using the “@” notation prefacing their username. For example, on the tweet “@cacobart imagine if it was raining coke, you’d love it” the author interacts with the user @cacobart.

We acknowledge four types of possible interactions via tweets: replies, retweets, mentions and attribution. A **reply** corresponds to a situation in which one user wants to answer a post from another one or simply direct the message to someone else. For example, a tweet of user  $A$  in reply to user  $B$  would be a post like “@B [content of the tweet]”. A **retweet** is used to propagate a message:  $A$  retweets  $B$  means that  $A$  posted a message that  $B$  has already posted. Retweets, particularly, either have a “RT” markup – for example, “RT @B [content posted by B]” – or have a Twitter official

retweet identification. Finally, a **mention** is a tweet that contains another user in the middle of the text (e.g. “[content] @A [content]”) and an **attribution** is similar to a retweet, except that it cites the username using the notation “(via @B)” instead of “RT @B”. We parse each gathered tweet and store all the interactions between users that discuss the topic for further analysis. Table 3.2 shows examples of each type of interaction.

### Complex Network Approach

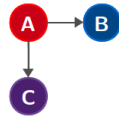
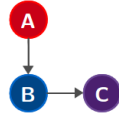
Next, in order to extract other metrics for characterizing the roles of users on Twitter and identify the influential ones, we adopt a complex network approach. From the several networks that naturally emerge from user relations enabled by Twitter features, we select two of them for an in-depth analysis: the Connection Graph ( $G_c$ ) and the Interaction Graph ( $G_i$ ). Formally, the networks are defined as follows.

**Definition 1 Connection Graph.** For a given subset of users involved in a specific theme, let  $(G_c, U)$  be the user directed unweighted graph, where  $(u_1, u_2)$  is a directed arc in  $U$  if user  $u_1 \in G_c$  follows user  $u_2 \in G_c$ .

**Definition 2 Interaction Graph.** For a given subset of users involved in a specific theme, let  $(G_i, U)$  be the user directed unweighted graph, where  $(u_1, u_2)$  is a directed arc in  $U$  if user  $u_1 \in G_i$  has cited at least once (i.e., mention, reply or re-tweet) user  $u_2 \in G_i$ .

Intuitively, the first network captures the declared connections between users (following-follower relation) whereas the second one captures the user interactions via tweets. Table 3.3 illustrates the construction of both graphs. Note that both graphs are unweighted, so the number of times a pair of users has interacted is not represented on them.

Table 3.3: Construction of Connection and Interaction Graphs.

Graph	Relation Between Users	Example
$G_c$	A follows B and C;	
$G_i$	A replies and mentions B in two different tweets; B retweets C;	

Different measures for networks analysis could be exploited (such as shortest paths, distance, component connectivity, clustering, clique, among others [Costa et al., 2007]). The measurements that make more sense for influence estimation are those based on centrality (defined on the vertices of a graph), because these metrics are designed to rank the notoriety of users according to their position in the network.

Similarly, influential users have to be well connected to other users, and play a central role in the graph in which they is embedded. For that matter, two centrality measures were chosen. We also analyze the in-degree of the users<sup>1</sup>, as follows.

- **Betweenness centrality (*bc*)** is the first centrality measure and is defined by the fraction of shortest paths between node pairs that pass through the node of interest [Brandes, 2008]. In both graphs  $G_i$  and  $G_c$ , users with high betweenness have an important role in the information dissemination process, since they act as bridges for the data flow.
- The centrality measure **Eigenvector centrality (*ec*)** [Bonacich, 2007, Ruhnau, 2000] considers that users are more central if they are related to users that are themselves central. Thus, the centrality of some node does not only depend on the number of its adjacent nodes, but also on their value of centrality. It is important to remark that Eigenvector centrality is an algorithm similar to Pagerank, applied to social networks [Chen et al., 2007]. We use this metric to rank higher users that are related to many other users or with a few users that are related to lots of other users.
- The **In-degree (*id*)** of each user is a key characteristic of the structure of a directed network. In the Interaction Graph, the in-degree measures the number of times a user was cited or had her tweets replied or retweeted, whereas in the Connection Graph, the in-degree stands out for the number of users within the topic that follows the user in focus.

It is important to emphasize that, in the Connection Graph, the *in* and *out*-degrees of the users are different from the following and follower counts that appear on their profile, because the degree concerns the connections between the users within the collected dataset.

---

<sup>1</sup>All metrics were calculated using `NetworkX` [Hagberg et al., 2010].

### Combined Metric for Relation Features

From an influence detection point of view, the most influential user for a topic is the one with the higher value for each of the four aforementioned metrics (*tf*, *bc*, *ec*, *id*). For this reason, the metrics were combined in an arithmetic mean (as shown in Equation 3.1).

$$u_{relation} = (bc + ec + id + tf)/4 \quad (3.1)$$

In order to combine the metrics equally, they were normalized individually to a  $[0, 1]$  scale [Jain, 1991]. Specifically, we did a *Range Normalization* [Jain, 1991], in which the range is changed from  $[x_{min}, x_{max}]$  to  $[0, 1]$ . The scaling formula is:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}},$$

where  $\{x_1, x_2, \dots, x_m\}$  are the measured values and  $x'_i$  the scaled value corresponding to  $x_i$ .

The result  $u_{relation}$  is also in this range. Due to the broad distribution of centrality measure values, the normalization of *ec* and *bc* was calculated using logarithmic quantities.

### 3.4.2 Polarity Features

The next perspective of our influence study corresponds to the author’s polarity. This sentiment analysis allows the detection of engagement of the users towards the defined topic and, consequently, leads to identifying users who are well connected regarding interactions and responsible for influencing others’ decisions due to the polarity of their tweets. Furthermore, in a “crisis management” point of view, recognizing the users who lead the positive and, mainly, the negative information flow is essential. The steps for extracting the user overall polarity on a given topic are described next.

#### Tweet Sentiment

In order to analyze the tweets as positive, neutral and negative, we first have to define the type of post that fits into each polarity. Primarily, **positive** tweets are the ones which promote the chosen topic, by expressing user appreciation or satisfaction. Likewise, **negative** ones express aversion towards the topic and may contain complaints, bad reviews, and so forth. **Neutral** tweets, on the other hand, are usually the ones that mention the topic in an unbiased way, for example, just mentioning it in another context or with a purely informative content.





Figure 3.3: Example of positive tweets about Paypal.

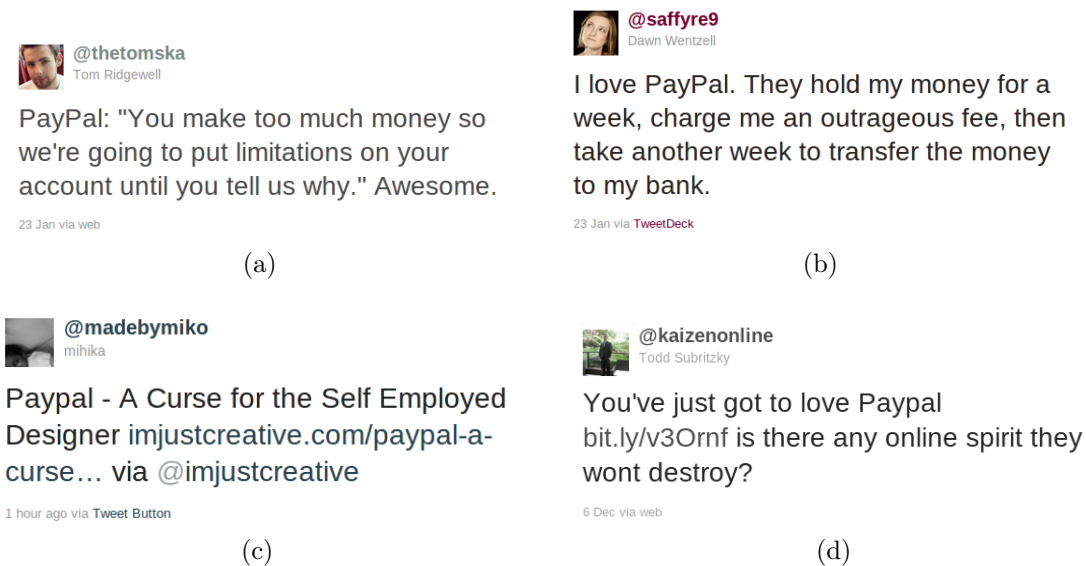


Figure 3.4: Example of negative tweets about Paypal.

Figures 3.3, 3.4 and 3.5 contain examples of tweets for each sentiment, concerning PayPal (an online service for payments and money transfers). These examples also emphasize the complexity of classifying tweets' sentiment. Aside from its short size, its content is often colloquial and filled with irony and sarcasm, both tones hard to identify. For instance, the negative tweets (a), (b) and (d) on Figure 3.4 have positive expressions, like “awesome”, “i love paypal” and “you’ve just got to love paypal” and, yet, in a pejorative way. Furthermore, the neutral tweet (d), in Figure 3.5, illustrates a tweet that seems like a negative one, but is neutral instead, because the negative opinion is towards ebooks.com, not Paypal.

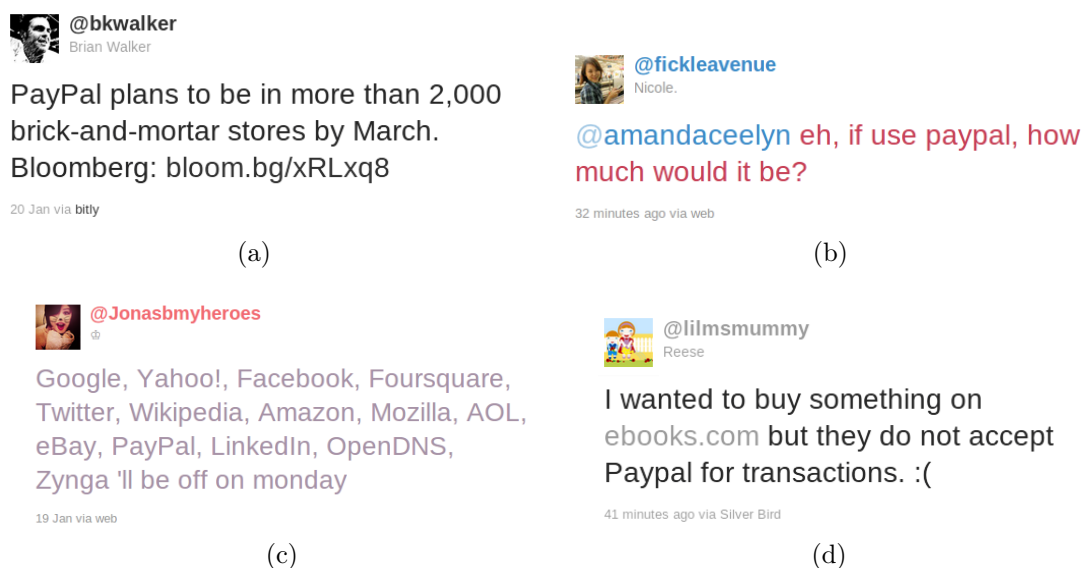


Figure 3.5: Example of neutral tweets about Paypal.

### Manual and Automatic Analysis

In this work, we employ two approaches for analyzing the tweets and evaluating our results: a manual and an automatic one. The goal to cover both types of analysis is to evaluate the potential of the influence detection method (using the manual analysis) and the impact of using automatic sentiment classification in comparison.

In the manual approach, tweets were manually classified by a marketing analysts' team, in a process in which each tweet's sentiment was verified at least by two analysts and a supervisor. In case of disagreement, the supervisor's decision was taken into account.

For the automatic approach, we employed a supervised classification, that is, a manual categorization is performed only for a set of tweets, used as training for a machine learning algorithm. The chosen algorithm was Support Vector Machine (SVM), a state-of-the-art classifier [Joachims, 1998, Vapnik, 1995].

### Combined Metric for Polarity Features

Based on the classification of tweets, we calculate the overall polarity of the users, i.e., their *overall* contribution to the topic discussion. If users post mostly positive-biased content, they are considered as potential evangelists. On the other hand, if they post mostly negative-biased content, they may be potential detractors. Users that do not have a biased content are considered neutral.

We consider that positive and negative tweets nullify each other. Thus, for each

user, the polarity value is the summation of the sentiment of all tweets of the user, as shown in Equation 3.2. In the formula,  $t_i$  is the  $i$ th tweet (of  $n_u$  total tweets) of user  $u$  and  $w_+$ ,  $w_o$  and  $w_-$  are the weights associated with positive, neutral and negative tweets, respectively. The weight is used for balancing the sentiments.

$$u_{polarity} = \sum_{i=1}^{i \leq n_u} t_i, \text{ where: } t_i = \begin{cases} w_+ & \text{if } t_i \text{ is positive} \\ w_o & \text{if } t_i \text{ is neutral} \\ w_- & \text{if } t_i \text{ is negative} \end{cases} \quad (3.2)$$

For instance, one may want to increase the weight of negative tweets to highlight detractors. Also, one may argue that if users made the effort to write a non-negative tweet on the topic, they are positively contributing to the spread of news about the subject, thus neutral and positive tweets could be considered as equivalent.

In this dissertation, following the specialists' instructions, we considered that there are three classes of tweet sentiment and that the neutral ones contribute (with lower intensity) to the user's positive polarity, by using the weights:

$$\begin{aligned} w_+ &= +2 \\ w_o &= +1 \\ w_- &= -2 \end{aligned}$$

Similarly to the network perspective, the polarity values were range normalized. Positive and negative values were normalized separately: positive values to  $[0, 1]$  and negative values to  $[-1, 0]$ . The normalization was calculated using logarithmic quantities.

### 3.4.3 Content Features

Finally, a lot can be extracted from the content a user publishes. In this perspective, we analyze the content of the tweet itself and what it may indicate about the user characteristics and intentions.

#### Tweet Readability

User-generated content is usually very heterogeneous, due to the variety of users' backgrounds and their different intentions. Our goal in analyzing the quality of the tweet content is to rank higher posts (and, consequently, their authors) that are well written and understandable. We hypothesize that if users are to influence other people, their tweets are expected to have a minimum quality.

For that matter, each tweet is evaluated using the Flesch-Kincaid Grade Level metric [Ressler, 1993] (*kincaid*), which was designed to indicate comprehension difficulty when reading a passage of contemporary academic English. This metric, successfully applied in the identification of high-quality Wikipedia articles [Dalip et.al, 2009], increased the accuracy of the influential identification for some cases, as studied in the experiments in Chapter 4.

For each tweet, it computes the average number of syllables per word and the average sentence length – see Equation 3.3. For instance, a tweet like “aaaaaaa haaate justin bieber!” has a low quality value, whereas “PayPal is dangerously easy.” a high one. Even though the readability metrics are not expected to work flawlessly for the short sized and noisy content of tweets, the results show that the metric helps eliminating undesirable content.

$$kincaid = 0.39 * \frac{words}{sentences} + 11.8 * \frac{syllables}{words} - 15.59 \quad (3.3)$$

For calculating this metric, we used the package `Style and Diction`<sup>2</sup>.

### Combined Metric for Content Features

The user quality perspective was determined as the average of the Kincaid metric computed for each tweet of the user, as defined in Equation 3.4.

$$u_{content} = \sum_{i=1}^{i \leq n_u} kincaid_i \times \frac{1}{n_u} \quad (3.4)$$

## 3.5 Influence Score

So far, we have presented different types of information that can help characterizing Twitter users, divided into three perspectives: relation, polarity and content. By exploiting them together, we aim to assign a single value (**influence score**) to each user in order to obtain a final (and possibly better) user rank.

The obvious intuition is to build a formula that combines the three feature sets. The user influence score ( $I_{score}$ ) is given by Equation 3.5, which, by the way, is one of the main contributions of this work.

$$I_{score} = \frac{\alpha * u_{polarity} + \varphi * (\beta * u_{relation} + \gamma * u_{content})}{\alpha + \beta + \gamma} \quad (3.5)$$

where:

---

<sup>2</sup>**Style and Diction Package:** <http://www.gnu.org/software/diction/diction.html>

$u_{polarity}$ ,  $u_{relation}$ ,  $u_{content}$  are the normalized polarity, relation and content perspectives;

$\alpha$ ,  $\beta$ ,  $\gamma$  are constants, greater or equal to zero, that weight each of the three perspectives; and

$$\varphi = \frac{u_{polarity}}{|u_{polarity}|} .$$

As mentioned before, both relation and content perspective values were normalized to fit into the range  $[0, 1]$ , whereas the polarity perspective values fit into  $[-1, 1]$ .

The auxiliary variable  $\varphi$  adjusts both relation and content perspectives according to the polarity result. If a user has a polarity equal to zero, the result of the equation is zero (regardless of the other features). Also, if the polarity is negative, both relation and content have their signal changed. The resulting influence score, for each user, is in the range  $[-1, 1]$ . By sorting the users in descending order, the top ones, with  $I_{score} > 0$ , are evangelists or neutral users and the bottom ones, with  $I_{score} < 0$ , detractors.

The idea behind combining different perspectives into a single influence score is that a feature alone may not be enough to characterize whether a user is influent or not, whereas the combination of the features may be. A user who is well positioned in the graph, has a biased opinion, and writes fairly well written tweets should be ranked higher as an influential user. The formula eliminates types of profiles that are erroneously appointed as influent. For example: (i) someone that is well connected to other users, but does not have a biased opinion about the subject; (ii) someone that posts, daily, hundreds of positive/negative tweets about the topic, but, for some reason, no one pays any attention to; (iii) a person whose content is too noisy and does not have a persuasive speech. For the specific cases listed above, the low values of polarity (i), relation (ii) and content (iii), respectively, do keep those users from being considered as influent.

## 3.6 Concluding Remarks

In this Chapter we have defined influential users for a topic as those responsible for producing an effect on other users. Particularly, their actions imply in other persons' actions; they act like bridges on the interactions about the topic; they have the intention of persuading the others positive or negatively; and they produce content with a minimum expected readability.

Based on this definition, we present SaID, our method for detecting influential users on Twitter. Our approach focuses on how users behave in a contextual topic

discussion. We try to characterize the user behavior by observing three facets: *relation*, *polarity* and *content*. In other words, we observe the polarity and quality of users' tweets and their position in two networks: (1) follower/following connections and (2) interactions via tweets. We extract several features from each user and combine them into an influence score. Finally, by ranking the users according to this score, we provide a list of evangelists and detractors for the specified topic.

# Chapter 4

## Experiments with Manual Assessment of Tweets' Sentiment

This chapter presents the experimental validation of SaID considering first a manual assesement of the sentiment of the tweets. A fully automatic approach is evaluated in Chapter 5.

We start this Chapter by describing the datasets built for influential detection purposes (Section 4.1). Then, we discuss statistics of the collections, the classification of their tweets and the influential users used as ground truth. Next, we list the metrics employed when evaluating SaID and details about the implemented baselines (Section 4.2). The following Sections discuss the actual experiments, divided into three parts. The first part compares the performance of SaID using the different graphs: Connection and Interaction (Section 4.3). We evaluate both approaches comparing their effectiveness on detecting the influential users. We also measure the execution time for calculating each network metric for both approaches. In the second part, we analyze the impact of each individual perspective (polarity, relation and content) on our method (Section 4.4). Furthermore, we present an approach for optimizing the Influence Factor of each user. Finally, the third part presents a comparison of SaID results against different baselines (Section 4.5).

### 4.1 Datasets

There is no established benchmark for evaluating user influence detection on Twitter. So, a major effort in this work was to build such datasets. Although expensive and demanding, this process is essential for the experimental validation presented in this Chapter. In the present Section, we describe each dataset and their characteristics.

### 4.1.1 Dataset Characteristics

We have built three collections for the experiments. The first one regards a *soda brand*, the second one regards a *home appliance brand* and the third regards a *groceries megastore chain*. Table 4.1 summarizes the content of each dataset. All the tweets are in Brazilian Portuguese and were posted by Brazilian users.

Table 4.1: Datasets’ characteristics.

Topic	Time interval	# users	# tweets
<i>soda brand</i>	August to September 2009 (1 month)	6,885	8,063
<i>appliance brand</i>	July to August 2010 (1 month)	1,617	2,354
<i>groceries megastore</i>	January 2nd to January 9th 2012 (1 week)	4,372	9,383

A simple analysis of the profile/tweet quantities makes it clear that the appliance brand topic is the least mentioned of the three datasets: in one month, only 1,617 users posted 2,354 tweets. By observing the tweets, one can see that the content of this dataset is mostly related to customer service conversations. The soda dataset has a larger amount of tweets and profiles when compared to the appliance dataset. The content of the soda dataset is very diverse. This type of product is also more present in the users’ daily activities than appliance brands, hence they are more mentioned in tweets. Finally, the groceries megastore dataset has the higher number of tweets (in the smallest interval of time), but a relatively smaller number of users. This behavior is interesting and reflects the content of the posts. On January 2nd, the groceries megastore chain announced on their online store several products on sale, but when the customers tried to buy the products, the purchases were cancelled. Several users complained about that fact on Twitter, while the groceries megastore official profile tried to fix the situation.

The three datasets are topic-related and concern specific brands. We have built these datasets with marketing scenarios in mind, in which users and what they say about the brand are to be monitored.

### Follower/Following Analysis

For comparing the collected (and topic-focused) datasets to the previously analyzed samples of the Twitter network [Huberman et al., 2008, Krishnamurthy et al., 2008, Kwak et al., 2010], we analyzed some follower/following statistics. The goal is to show the similarities between topic-focused datasets and general ones.



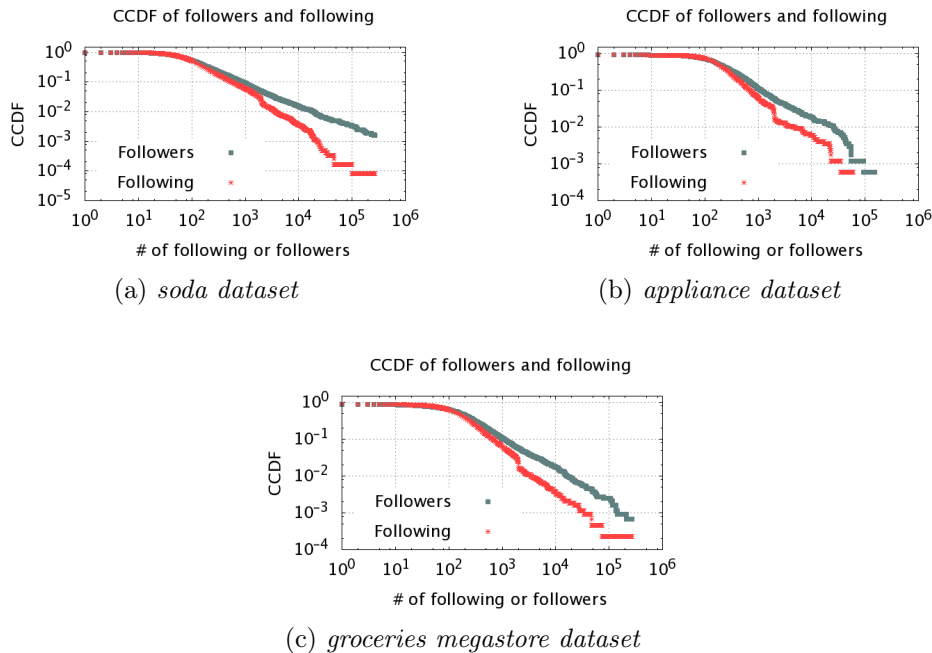


Figure 4.1: CCDF probabilities of following and followers for each dataset.

Specifically, we first analyzed the distribution of following and followers in a complementary cumulative distribution function (CCDF). In statistics and probability theory, CCDF describes the probability of a given value  $X$  taking a value above a particular level [Jain, 1991]. That is,  $\bar{F}(x) = P(X > x)$ . The y-axis of the plots in Figure 4.1 represents the CCDF probability. The blue square points represent the number of follower relations whereas the red asterisks represent following relations of a user for the specified dataset.

These distributions, specially the region beyond  $x = 10^4$ , on the three datasets, have a similar behavior to the one reported by Kwak *et. al* in [Kwak et al., 2010]. This “stair-like behavior” shows that there is a lack of users that follow and are followed by more than  $10^4$  profiles.

Still observing the follower/following relations, the plots in Figure 4.2 shows the  $\frac{\# \text{ follower}}{\# \text{ following}}$  ratio distribution among the users for each dataset. It is possible to identify in these plots each type of user, according to the aforementioned *Twitter Follower-Followee Ratio*: high ratio users ( $\uparrow$  followers,  $\downarrow$  following) appear in the region above the diagonal; users with ratio approximately 1 (followers  $\simeq$  followees) are around the  $y = x$  line; and users whose ratio approaches zero ( $\downarrow$  followers,  $\uparrow$  followees) are located below the diagonal. By comparing the *tff* plots with previous work, such as [Krishnamurthy et al., 2008], there are fewer representatives of the last group in all three datasets. Since

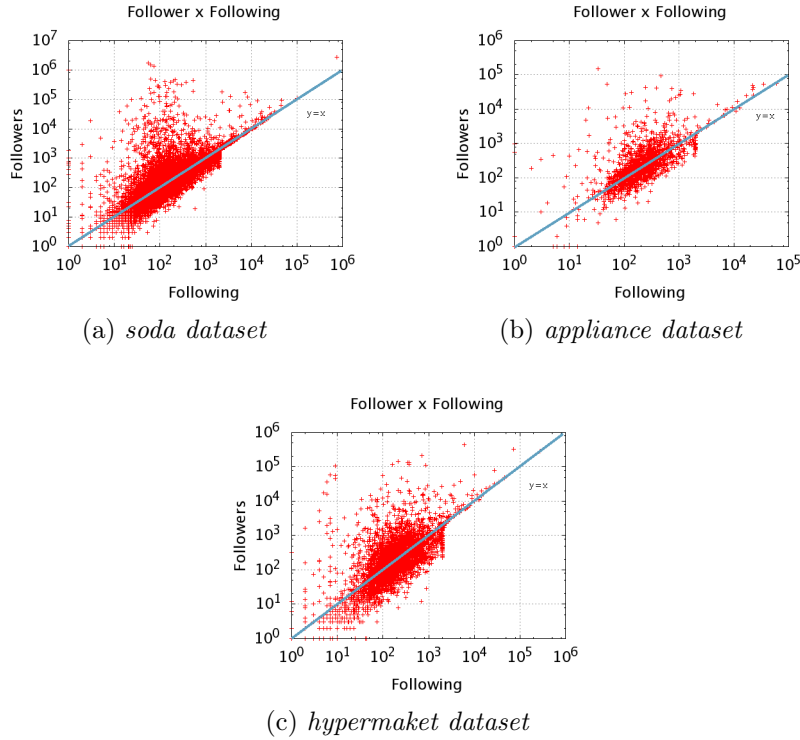


Figure 4.2: The relation between the number of followers and following of a user.

the tweets posted by the third group are usually classified as noise (they may contain the keywords but often have unrelated advertising associated) and the set of users is built from the posted tweets, their representation in this dataset is smaller than usual. In order to be an influential user, the person must be an author and must tweet.

The similarities between the subject-restricted dataset and the other generic samples of Twitter show that there are correspondent types of user in both contexts, which indicates that our method can be expanded to a wider context.

### 4.1.2 Tweet Sentiment Classification

According to the methodology for sentiment analysis (described in Section 3.4.2), the tweets of each dataset were classified as positive, negative, neutral or noise (for tweets that do not correspond to the respective topic).

For classifying the polarity of the tweets, we first perform a manual analysis and, later, an automatic analysis, contrasting both. The influence scores calculated from the manual analysis of the tweets represent a skyline for our method and are used in the experiments conducted in Sections 4.3, 4.4 and 4.5. Our goal in automatically analyzing the content is to verify the effectiveness loss while using a fully automatic

method (Section 5). The manual classification of the tweets employed for most of the experiments here is described below.

### Manual Analysis

The manual classification was performed by a team of specialists in marketing and communication. The specialists responsible for the tweet’s classification are native speakers of Brazilian Portuguese (the datasets’ language). Table 4.2 presents the number of tweets and users for the datasets along with the respective sentiment classification.

The soda and the groceries megastore datasets have a majority of neutral tweets, whereas the appliance one has a majority of positive. Soda brands are more present in people’s routine than appliance brands. That is, soda brands may be cited in tweets that do not specifically focus on soda (for example, if a user says that she’s drinking coke while having dinner). This does not happen so frequently with appliance brands and, for that reason, tweets tend to be more polarized.

The groceries megastore chain dataset has a peculiar behavior. It is, as the soda brands dataset, a topic present in people’s routine (most of the tweets mention the groceries megastore chain occasionally, for example, coming or going to a store), which may be the reason for the larger amount of neutral tweets. However, during the week in which we collected the tweets, users had problems while purchasing items on sale in the groceries megastore online store. This fact generated a huge amount of negative tweets.

Observing the tweets per user distribution, in Figure 4.3, we can see that the three datasets have similar distributions. In all datasets, the majority of the users

Table 4.2: Tweets and users per sentiment.

	Positive	Neutral	Negative	Total
<i>soda</i>				
<b>tweets</b>	3,083 (38.23%)	4,156 (51.54%)	824 (10.21%)	8,063
<b>users</b>	2,770 (40.23%)	3,401 (49.39%)	714 (10.37%)	6,885
<i>appliance</i>				
<b>tweets</b>	1,489 (63.25%)	580 (24.63%)	285 (12.10%)	2,354
<b>users</b>	1,198 (70.18%)	360 (21.08%)	149 (8.72%)	1,707
<i>groceries megastore</i>				
<b>tweets</b>	132 (1.41%)	5,589 (59.56%)	3,663 (39.03%)	9,383
<b>users</b>	109 (2.49%)	2,678 (61.25%)	1,582 (36.18%)	4,372

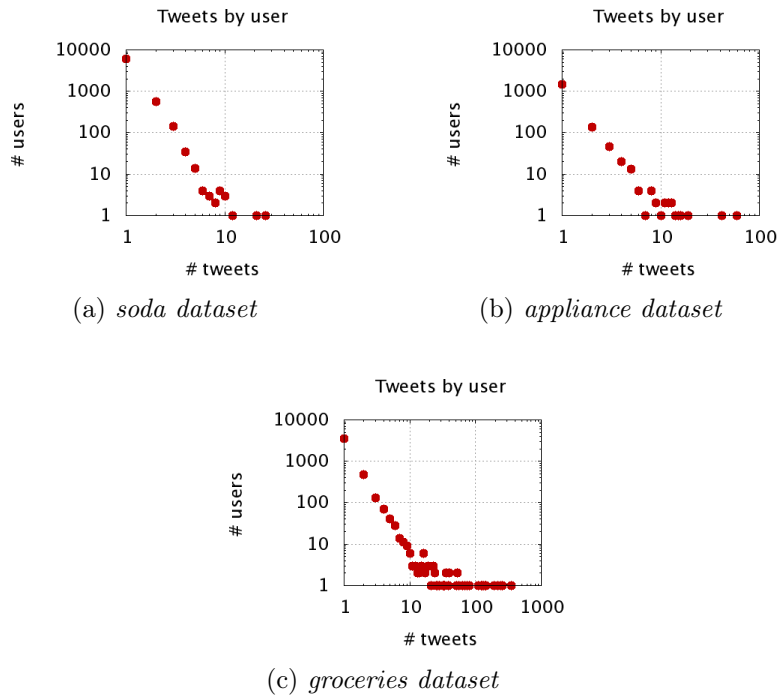


Figure 4.3: Quantity of tweets posted by users.

posted less than 10 tweets. However, for the groceries megastore dataset, the scenario in which *a few users post a large amount of tweets* occurs more often. The user with the highest number of tweets in this dataset, posted 352 tweets complaining about his problems while purchasing a product.

Finally, this dataset considers all users that posted a message regarding the subject. In other words, the dataset considers users with regular frequency of tweets, users with only one tweet, and even users with a high number of tweets. All those cases may concern an influential users, depending on the users' characteristics and the repercussion of their posts.

### 4.1.3 Influential Users: Ground Truth

Finally, for testing SaID, we needed a ground truth list of who was indeed influential for the collected datasets. For two of the datasets (soda and appliance) the specialists' team produced lists of evangelists and detractors. For the groceries megastore chain dataset, we created an evaluation pool, in which participants rated the influence of users. Both methods (and their results) are discussed next.

### Specialists' Analysis

For the two first datasets (soda and appliance), the same marketing and communication specialists team that analyzed the sentiment of the tweets created a list of influential users for the datasets.

The procedure was analogous to the one for sentiment classification: at least two analysts classified each user as influent or not, and a supervisor checked the results, handling the disagreements. The claimed intuition was that users whose content was widespread, whose tweets were engaged towards a point of view and whose importance among the topic was relevant, were influential. They analyzed information about the tweets (RTs, replies) and the users (who they are, what type of tweets they usually write, what the repercussion of their tweets was and so on).

For the soda dataset, they found **17 influential users**: 10 *evangelists* and 7 *detractors*. Meanwhile, for the appliance dataset, they found **39 influential users**: 23 *evangelists* and 16 *detractors*.

Although the quantity of users found to be influent seems small, no limit was imposed to the analysts in terms of maximum number of influential users per data set. The team is used to this type of analysis and usually provides such service commercially.

### Evaluation Pool

The last dataset, the one about an groceries megastore chain, had its users' influence differently evaluated on a user study. An evaluation pool was constructed and populated with the top users from diverse baselines<sup>1</sup> as well as ranks generated by SaID.

Specifically, we have used the top 10 evangelists and top 10 detractors found by each of the following methods:

1. **Klout Baseline** (KB): 10 positive and 10 negative users with highest Klout Score;
2. **Tweet Baseline** (TB): 10 positive and 10 negative users with highest tweet count;
3. **Follower Baseline** (FB): 10 pos. and 10 neg. users with highest follower count;
4. **Random Baseline** (RB): one random list for positive and one for negative users;
5. **SaID**, using only the polarity features;
6. **SaID**, using only the relation features;
7. **SaID**, using only the content features.

---

<sup>1</sup>The baselines are described in detail, in Section 4.2.2.

As expected, there was some overlap among the ranked lists generated by the 7 methods. Eliminating the duplicates, we obtained 100 users who would have their influence evaluated.

Most of the participants of the pool were graduate students in Computer Science. Only two had a marketing and communication background. Each participant evaluated 25 users randomly selected from the 100 users in the evaluation set. 21 participants helped on this influence survey and each user in the dataset was evaluated by 5 different participants.

The users were presented for evaluation along with their login, profile on Twitter, tweets on the dataset (with links to the original content on Twitter) as well as their number of interactions via tweets (mentions, replies, RTs, attributions). The participants were instructed to find users whose content was widespread, whose tweets were engaged towards a point of view and who were important among the topic.

Based on this information, the participant had to answer if each user had “*High Influence*” (HI), “*Low Influence*” (LI) or “*No Influence*” (NI) at all, considering the topic. Additionally, we gave the participants the option to mark if they are "uncertain" about their analysis. Figure 4.4 shows an example of a user shown to a participant.

In order to assess the reliability of agreement between the raters when assigning the users' categories, we employ Fleiss' kappa statistical measure [Fleiss, 1971]. This measure gives the degree of agreement  $\kappa$  of the raters (Equation 4.1) when a fixed number of people assign categorical (or numeric) ratings to a number of items.

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \quad (4.1)$$

In Equation 4.1, the numerator  $\overline{P} - \overline{P}_e$  gives the degree of agreement *achieved* above chance and the denominator  $1 - \overline{P}_e$  gives the degree of agreement *that is attainable* above chance. Let  $N$  be the number of subjects,  $n$  be the number of ratings per subject

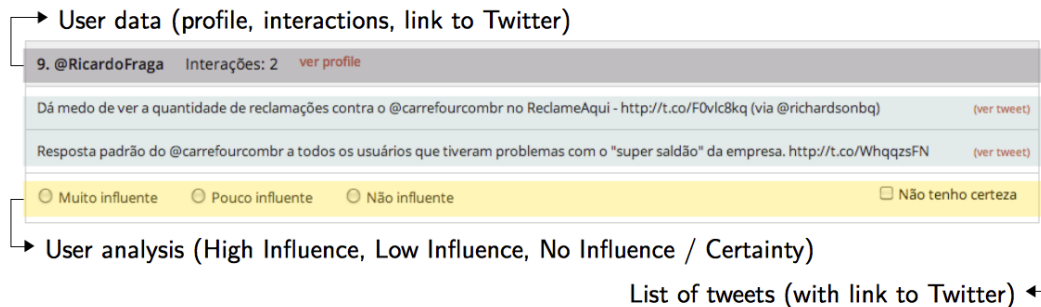


Figure 4.4: User data and analysis.

and  $k$  the number of categories into which assignments are made. The subjects are indexed as  $i = 1, \dots, N$  and the categories as  $j = 1, \dots, k$ . So,  $n_{ij}$  represents the number of raters who assigned the category  $j$  to the subject  $i$ .

This way,  $p_j$  is the proportion of all assignments to the  $j$ -th category (Equation 4.2a) and  $P_i$  is the level of agreement for the  $i$ -th subject, that is, how many rater-rater pairs are in agreement, relative to the number of all possible rater-rater pairs (Equation 4.2b). Finally,  $\bar{P}$  and  $\bar{P}_e$  are defined in Equation 4.2c.

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij} \quad (4.2a)$$

$$P_i = \frac{1}{n(n-1)} \left[ \left( \sum_{j=1}^k n_{ij}^2 \right) - (n) \right] \quad (4.2b)$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i \text{ and } \bar{P}_e = \sum_{j=1}^k p_j^2 \quad (4.2c)$$

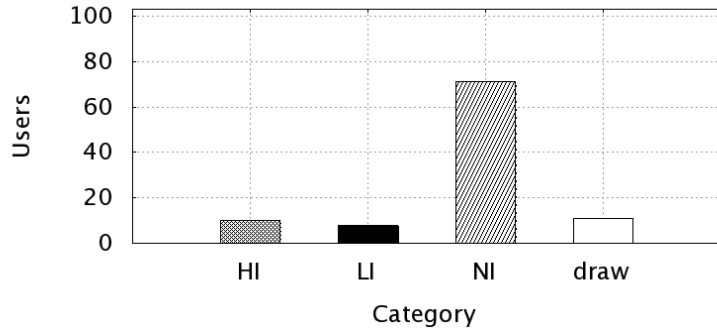
When the raters are in complete agreement,  $\kappa = 1$ . If there is no agreement among the raters, then  $\kappa \leq 0$ .

We considered two scenarios when categorizing the users. For the first one, with the three defined categories (HI, LI and NI), as shown in Figure 4.5a, we found 10 users classified as HI, 8 users as LI and 71 users as NI. However, 11 users did not have a category defined. Let  $n_{LI}$  be the number of raters that categorized a user as LI,  $n_{HI}$  as HI and  $n_{NI}$  as NI. These 11 users had either  $n_{HI} = n_{LI} > n_{NI}$ ,  $n_{HI} = n_{NI} > n_{LI}$  or  $n_{LI} = n_{NI} > n_{HI}$ . We represent these cases in the *draw* column.

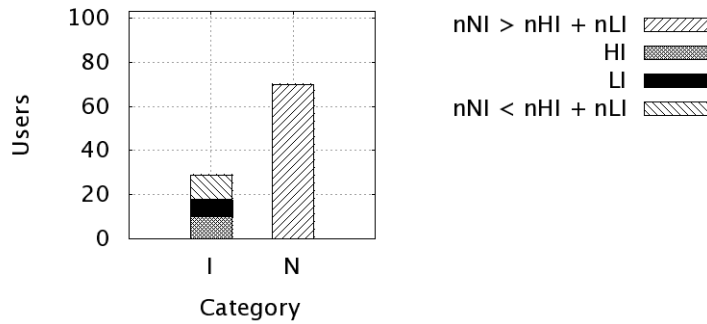
In the second scenario, we considered only two categories: *Influential* (I) and *Not Influential* (N). In this case, the rates for both HI and LI categories were considered as one. Figure 4.5b show the results in details: 70 users remained as N and 30 users were categorized as I.

Considering the three categories, the inter-rater agreement between the participants was  $\kappa = 0.18$ , which is considered *slight agreement*. However, considering only two categories the inter-rater agreement between the participants was  $\kappa = 0.24$ , which is considered *fair agreement* [Fleiss, 1971].

Classifying users as influential or not is a very subjective task. Even though instructions were given about what is considered influential, the participants opinions vary as to whether a user is influential or not. Moreover, to differ what is "high" and "low" influence is even more subjective. Thus, a higher agreement is expected when the classification occurs considering only two categories.



(a) Preliminary pool results, in which users were divided in the categories *High Influence* (HI), *Low Influence* (LI) and *No Influence* (NI). The users whose category could not be decided are represented in the column *draw*.



(b) Final pool results, in which users were divided in the categories *Influential* (I) and *Not Influential* (N). Specifically, the plot shows the new categories in terms of the preliminary ones. The Influential group contains users primarily categorized as HI, LI or users whose previous sum of classifications  $n_{HI} + n_{LI}$  is greater than their  $n_{NI}$  classifications. The Not Influential group contains users whose previous sum of classifications  $n_{HI} + n_{LI}$  is lower than their  $n_{NI}$  classifications.

Figure 4.5: Evaluation pool results.

Conclusively, we divided the 30 influential users into evangelists and detractors, according to their polarity, as seen next.

### Summarizing the Results

Table 4.3 summarizes the influential users found in each dataset, categorized as evangelists and detractors. The datasets have very different types of users and behaviors, which obviously affects the number of influential users. In the next Sections, we perform several experiments using these users as ground truth.

It is important, for future reference, to note the difference in the proportions  $\frac{\# \text{ influential}}{\# \text{ users}}$  for each sentiment and dataset. For the groceries megastore, evangelists



Table 4.3: Number of influential users (evangelists and detractors) in each dataset and the fraction that they represent.

	<i>soda</i>	<i>appliance</i>	<i>groceries megastore</i>
<b>evangelists</b>	10 (0.36% of + users)	23 (1.91% of + users)	8 (7.33% of + users)
<b>detractors</b>	7 (0.98% of − users)	16 (10.73% of − users)	22 (1.39% of − users)
<b>total</b>	17 (0.24% of the users)	33 (1.93% of the users)	30 (0.75% of the users)

represent around 7% of the positive users. A similar situation happens with detractors in the appliance dataset, with 10%. The lowest proportion of influential among total users is for evangelists in the soda dataset. As can be seen in the future experiments, this category is harder to find in this dataset.

## 4.2 Experiment Setup

Before moving on to the actual experiments, we describe our experiment setup, explaining the evaluation metrics (Section 4.2.1) and the baselines employed to evaluate our approach (Section 4.2.2).

### 4.2.1 Evaluation Metrics

In order to evaluate our method, we employ ranking performance measures [Baeza-Yates and Ribeiro-Neto, 1999], assuming the ground truth influential lists described in Section 4.1.3.

The measures *precision* and *recall* were adjusted to the context of detecting influential users, as shown in Equations 4.3 and 4.4, in which  $n_r$ ,  $n_{ir}$  and  $n_{it}$  are: the number of users in the method’s ranked list, the number of influential users in the method’s ranked list and the total number of influential users in the dataset.

$$precision = \frac{n_{ir}}{n_r} \quad (4.3)$$

$$recall = \frac{n_{ir}}{n_{it}} \quad (4.4)$$

Based on these two measures, we calculate the *F-score*,  $\mathcal{F}_\beta$ , of each rank as defined by Equation 4.5. This measure can be interpreted as a weighted average of precision

and recall.

$$\mathcal{F}_\beta = (1 + \beta^2) \times \frac{\textit{precision} \times \textit{recall}}{(\beta^2 \times \textit{precision}) + \textit{recall}} \quad (4.5)$$

SaID was designed to assist social analysts on the monitoring task by providing a list of TOP- $x$  evangelists and detractors. As a manner of measuring its quality according to the ranked list size available, we evaluate our results using what we call  $[measure]@x$ , meaning the measure (precision, recall or  $\mathcal{F}_\beta$ ) value at a user ranked list of size  $x$ . We use the notations  $\textit{recall}@x$ ,  $\textit{precision}@x$  and  $\mathcal{F}_\beta@x$ .

We have decided to evaluate rank lists of size  $x$ , considering  $10 \leq x \leq 150$ . This interval was chosen based on the specialists' team reasoning about an acceptable list size for analyzing possible influential users.

The earliest (the shortest ranked list size) the method reaches the measures' maximum value, the higher is its performance. Therefore, our goal is to optimize each  $[measure]@x$  **curve**, considering  $10 \leq x \leq 150$ . We evaluate this, by calculating the area below the measure curve, for which we use the notation  $a([measure])$ . That is,  $a(\textit{recall})$ ,  $a(\textit{precision})$  and  $a(\mathcal{F}_\beta)$ . Figure 4.6 shows, as an example, the best possible  $\textit{recall}@x$  curve.

As claimed by the specialists, the number of influential users in a dataset is usually small when compared to the total of users. Due to this fact, although high precision is desired, it is far more valuable to evaluate whether the method is able to find all the influential users or not. For that matter, we focus on maximizing  $\textit{recall}@x$ . Also, we employ  $\beta = 2$  in our  $\mathcal{F}_\beta$  evaluations ( $\mathcal{F}_2$  weights recall higher than precision).

Finally, the superscript notation  $i = \{e, d\}$  is used in some metrics to denote experiments/results on evangelists or detractors, specifically.

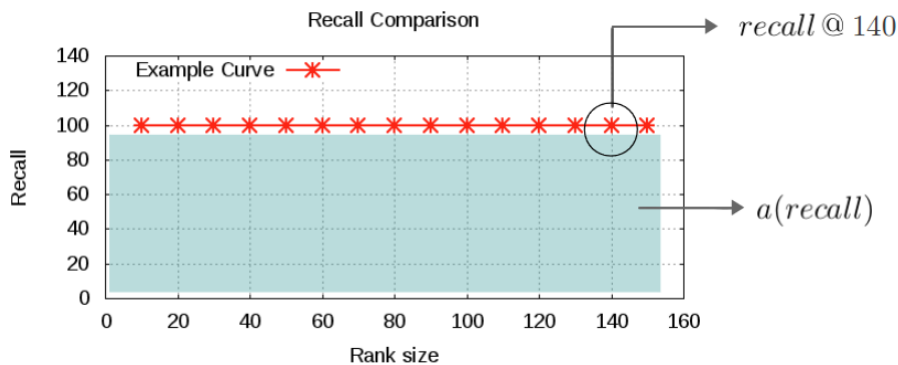


Figure 4.6: Example of plot for  $\textit{recall}@x$ ,  $10 \leq x \leq 150$ .

## 4.2.2 Baselines

Four baselines were implemented for evaluating SaID. In this section we define and evaluate the performance of each one of them.

### Klout

The first baseline used to evaluate our method is based on Klout<sup>2</sup>. Klout is a method for detecting influential users widely used on the Web and based on the users' “*ability to drive action*”. The Klout Score measures influence by analyzing three key aspects: true reach – how many people the user influence –; amplification – how much the user influence them –; and network impact – the influence of the user network.

We use Klout API<sup>3</sup> to retrieve the influence metrics returned by Klout for each user in our datasets. In order to build our Klout Baseline (KB), we rank the top  $x$  **positive** users with highest Klout Score as *evangelists* and the top  $x$  **negative** users with highest Klout Score as *detractors*.

### Tweet Baseline

Next, we created what we call Tweet Baseline (TB), in which we divide the users by polarity and order them by the number of tweets posted on the topic. That is, we rank *evangelists* as the **positive** users with the highest number of tweets and *detractors* as the **negative** users with the highest number of tweets.

The intuition of this baseline is to focus on users that contribute the most to the conversations about the topic, in terms of quantity of tweets. The baseline is slightly more intelligent than just ordering the users by amount of tweets, because we consider the polarity of the user. Therefore, official profiles would not appear on the top of the list. For example, profiles like *Coke*, in the soda dataset, *Brastemp* in the appliance dataset, and *Carrefour* in the groceries megastore dataset.

### Follower Baseline

Another baseline used for evaluating our method was the Follower Baseline (FB). In this baseline, the users were also divided by polarity and ranked according to their number of followers. That means that *evangelists* are the **positive** users with highest number of followers and *detractors* are the **negative** users with higher number of followers.

---

<sup>2</sup><http://klout.com/>

<sup>3</sup>[http://developer.klout.com/api\\_gallery](http://developer.klout.com/api_gallery)

Our goal, by building this baseline, is to explore the number of followers of a user as an indicator of influence, in other words, the higher the audience, the higher the influence.

### Random Baseline

The last baseline employed in our experiments is the Random Baseline, RB, in which two *random* lists of users are generated: one for positive users and one for negative users.

### 4.2.3 Analysis of the Baselines

With the ground truth influential users at hand, we evaluate the performance of each of the aforementioned baselines, by presenting *recall @ x* plots, as follows.

#### A Note on the Random Baseline Values

The measures *recall @ x* and  $\mathcal{F}_2 @ x$  presented for the Random Baseline (RB) in the following Sections were calculated as the mean of several samples. The steps for calculating these values are explained below.

Firstly, for each dataset (*soda*, *appliance* and *groceries megastore*) and polarity (evangelists and detractors) we sampled 100 different random sets of users and found the values for the measures *recall* and  $\mathcal{F}_2$  for each rank size. We represent these values as  $y_1^i(m @ x)$ ,  $y_2^i(m @ x)$ ,  $\dots$ ,  $y_{100}^i(m @ x)$ , where  $m$  represents each measure  $m = \{\textit{recall}, \mathcal{F}_2\}$ ,  $x$  represents each rank size  $10 \leq x \leq 150$  and  $i$  represents evangelists or detractors  $i = \{e, d\}$ .

For every combination (dataset - polarity - metric - rank size), we calculated the sample mean  $\overline{y^i(m @ x)}$  of the 100 repetitions as well as the sample standard deviation  $s^i(m @ x)$ . In Table 4.4 we present the notation used, in details.

Based on these values, we calculated  $n^i(m @ x)$ , as the minimal number of observations required for the measure  $m$ , at rank size  $x$ , to provide an accuracy of  $\pm 20\%$ , with a confidence level of 80%. Equation 4.6 shows how to calculate this number of observations, according to Jain [1991]. In the Equation,  $r$  is the desired accuracy and  $z$  is the normal variate of the desired confidence level, that is,  $r = 20$  and  $z_{0.9} = 1.282$ .

$$n^i(m @ x) = \left( \frac{100z s^i(m @ x)}{r \overline{y^i(m @ x)}} \right)^2 \quad (4.6)$$

Table 4.4: Notation used for  $recall @ x$  and  $\mathcal{F}_2 @ x$  measures  $m$  calculated for the polarity  $i = \{e, d\}$  for each dataset.

		$x = 10$	$x = 20$	$\dots$	$x = 150$
samples	<b>1</b>	$y_1^i(m @ 10)$	$y_1^i(m @ 20)$	$\dots$	$y_1^i(m @ 150)$
	<b>2</b>	$y_2^i(m @ 10)$	$y_2^i(m @ 20)$	$\dots$	$y_2^i(m @ 150)$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	<b>100</b>	$y_{100}^i(m @ 10)$	$y_{100}^i(m @ 20)$	$\dots$	$y_{100}^i(m @ 150)$
sample mean	$\overline{y^i(m @ 10)}$	$\overline{y^i(m @ 20)}$	$\dots$	$\overline{y^i(m @ 150)}$	
standard deviation	$s^i(m @ 10)$	$s^i(m @ 20)$	$\dots$	$s^i(m @ 150)$	
observations needed	$n^i(m @ 10)$	$n^i(m @ 20)$	$\dots$	$n^i(m @ 150)$	

Finally, we calculated a single number of observations  $n_m$  for each dataset and metric  $m$ , as the maximum  $n^i(m @ x)$ , considering all rank sizes and polarities, as shows Equation 4.7.

$$n_m = \max(n^i(m @ x)), 0 \leq x \leq 150 \text{ and } i = \{e, d\} \quad (4.7)$$

Table 4.5 summarizes the number of observations needed for each dataset and metric. The high number of repetitions needed is a consequence of the small number of influential users. For example, considering the soda dataset, one influential accounts for 5.88% of the influential users set (1/17), leading to a high standard deviation, and consequently to a large number of samples needed for the given confidence and error.

Table 4.5: Number of observations needed for  $recall$  and  $\mathcal{F}_2$  random plots, with 20% of accuracy and a level of confidence of 80%.

	<i>soda</i>	<i>appliance</i>	<i>groceries megastore</i>
<i>recall</i>	6931	6331	12011
$\mathcal{F}_2$	975	2782	5000

## Comparing the Results

Figure 4.7 shows the evangelist and detractor’s plots, for each one of the baselines, for the soda dataset. Of the four baselines, the one with worst result, as expected, is the random one. As discussed earlier, in this dataset, the proportion between influential users and the total of users is very low (0.39% for evangelists and 0.98% for detractors). This fact is directly reflected on the random curve.

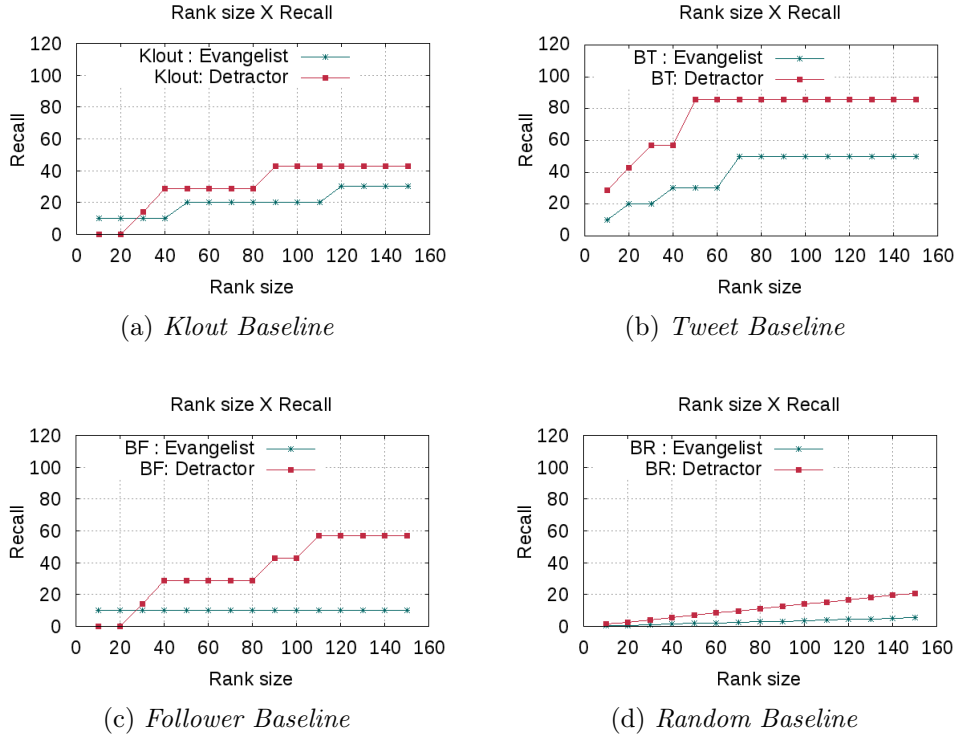


Figure 4.7: Values of  $recall @ x$  using each baseline, for *soda* dataset.

Still in the *soda* dataset, for all the baselines, finding the evangelists is harder than finding the detractors. Besides the fact that there are less evangelists among the positive users than detractors among negative users, the positive ones are harder to identify and are often misclassified as neutral users. Moreover, it is important to notice that no baseline reaches 100% of recall neither for evangelists nor for detractors for this dataset.

In the *appliance* dataset (Figure 4.8) the baselines have a behavior similar to the *soda* dataset: detractors are easier to find. As expected, by observing the proportion of detractor users in this dataset, the negative curve for the Random Baseline is much better than the positive. Both Random and Follower Baselines reach 100% of recall on the rank list size  $x = 150$ . The Tweet Baseline is the fastest one, reaching 100% of recall at  $x = 120$ .

Finally, the plots of the *groceries megastore* dataset in Figure 4.9 demonstrate the difference between the selection of influential users by specialists in marketing and communication (*soda* and *appliance* dataset) and regular people (on the evaluation pool for this dataset). The number of followers seemed to be a relevant factor for the participants of the pool, when classifying the users as influential or not, because both

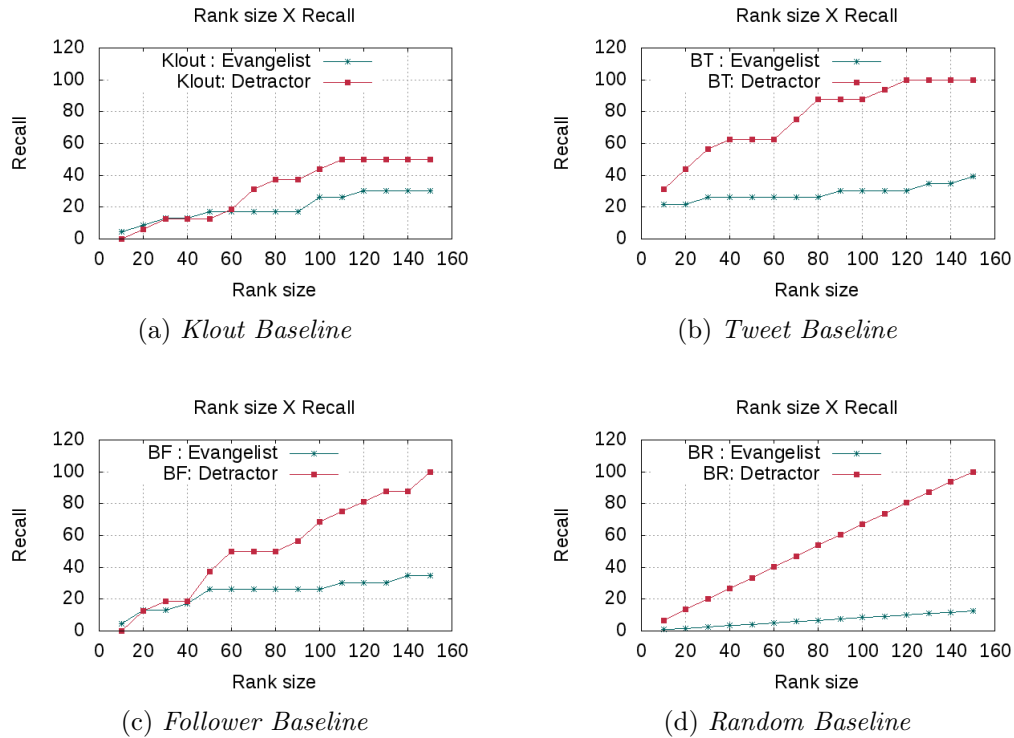


Figure 4.8: Values of  $recall @ x$  using each baseline, for *appliance* dataset.

Klout and Follower Baselines had their best results in this dataset. For the other two datasets, the content of the tweets and polarity of the users (most captured in the Tweet Baseline) were more important, as the better performance in these cases points out.

Furthermore, in this dataset, due to the small number of positive users (only 109) and the nature of the baselines (all of them provide positive and negative rank lists of users according to an ordering metric), finding the evangelists is easier. Specifically, after the rank size  $x = 100$ , all the baselines reach 100%. Finally, as expected, in the Random Baseline, the evangelist curve is better than the detractor's (due to the difference of proportions).

### 4.3 Interaction $\times$ Connection Approaches

The first part of the experiments conducted to validate SaID is a comparative analysis between the two proposed approaches for finding influential users: one using the Connection and other using the Interaction Graph.

This Section presents the results for influence detection in each dataset using the

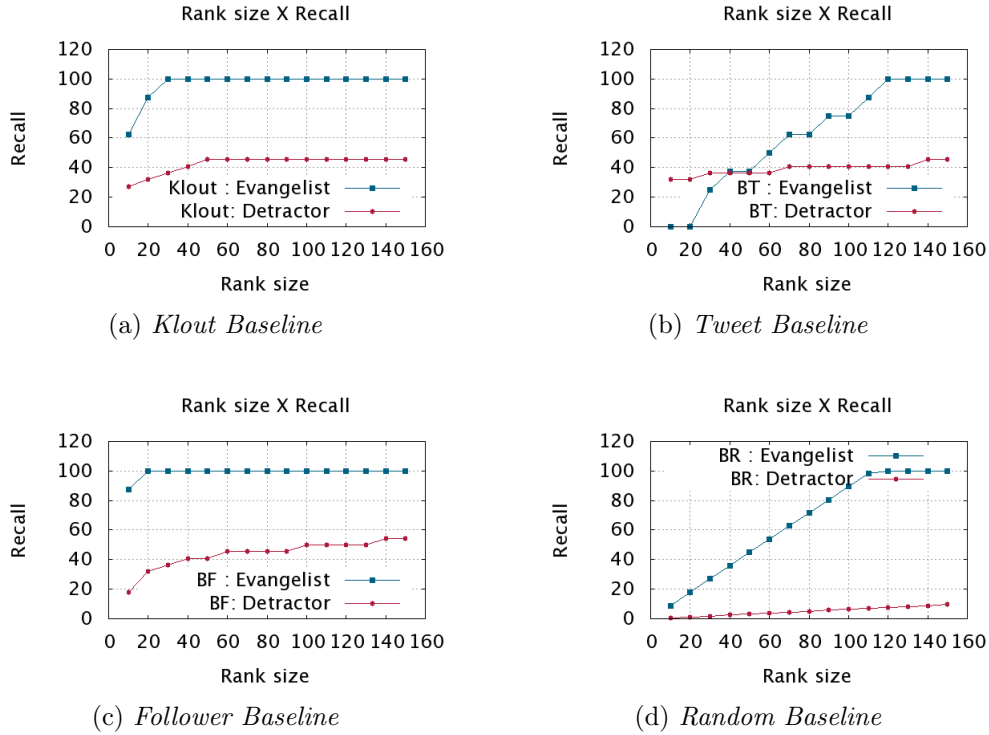


Figure 4.9: Values of  $recall @ x$  using each baseline, for *groceries megastore* dataset.

manual sentiment analysis.

### 4.3.1 Plotting the Graphs

Both Connection  $G_c$  and Interaction  $G_i$  Graphs were constructed for each dataset. Table 4.6 compares the number of vertices and arcs of both graphs  $G_i$  and  $G_c$  built based on *soda*, *appliance* and *groceries megastore* dataset and Figure 4.10 displays a visual representation of them.

As shown by Huberman et al. [2008] (and visible in Figure 4.10), the graph of interaction is considerably more sparse than the connection graph for all datasets.

Table 4.6: Statistics for  $G_i$  and  $G_c$  for each dataset.

	Nodes	Arcs in $G_i$	Arcs in $G_c$
<i>soda</i>	6885	797	8473
<i>appliance</i>	1707	1009	6103
<i>groceries megastore</i>	4372	2755	3549



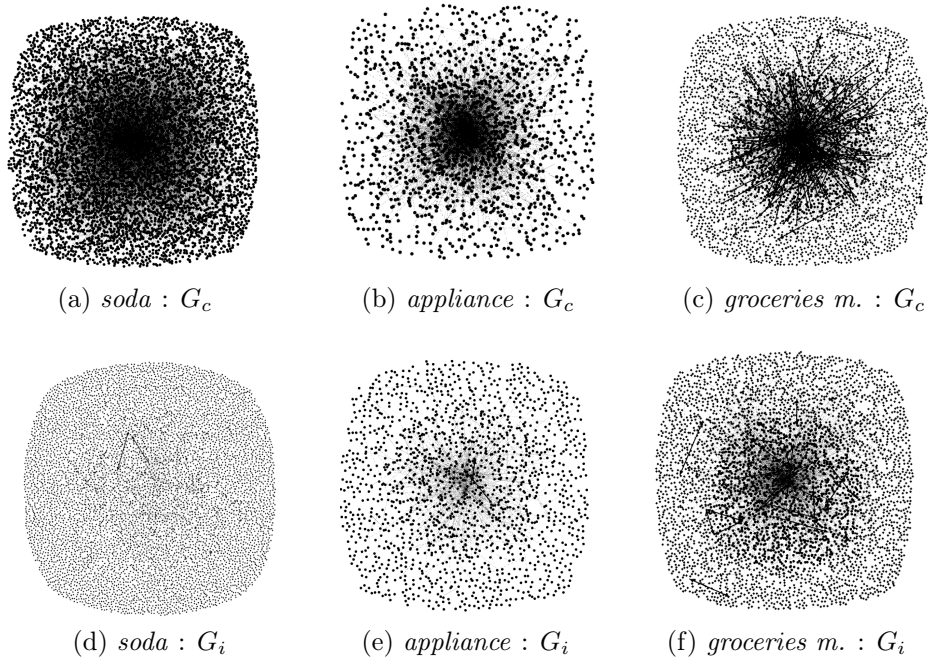


Figure 4.10: Graphic representation of  $G_i$  and  $G_c$  for the datasets.

Accordingly, the number of arcs in  $G_c$  is much larger than in  $G_i$  in the three cases. This fact is coherent with the fact that users interact with fewer people when compared to the ones they follow or are followed by.

Due to the velocity with which users and connections change in the online environment, there may be arcs not represented in  $G_c$ , due to changes in the user profile. Users may change their usernames or protect their accounts during the experiments, making it unavailable to collect their data. We expect these changes to be not significant, though.

### 4.3.2 Interaction $\times$ Connection-based Influence Detection

For comparing the approaches, two types of influential users ranked lists were generated for each dataset: one using the Interaction Graph ( $G_i$ ) as source for the relation features and the other using the Connection Graph ( $G_c$ ).

In order to calculate the users' Influence Score using the formula on Equation 3.5, a parameter combination  $(\alpha, \beta, \gamma)$  must be defined. Section 4.4 discusses in detail the problem of parametrization of the formula. Meanwhile, for the experiments in this section, we calculated the influence factor of each user (and approach) using a *leave-one-out* procedure, in which the parameters  $\alpha$ ,  $\beta$  and  $\gamma$  for each user's Influence Score are learned. This procedure is explained in details, in Section 4.4.

### Overall Performance

Firstly, for each approach, we determined its overall performance by measuring the  $\mathcal{F}_2$  of the ranks generated. Table 4.7 shows the  $\mathcal{F}_2^i$ ,  $i = \{i, e\}$ , values for the generated ranked lists. The absolute values are calculated at ranked lists of size  $x = 150$ . The area values  $a(\mathcal{F}_2^i)$  are calculated for  $10 \leq x \leq 150$ .

For the *soda* dataset, all the values for Interaction Graph are higher than the ones for Connection Graph. For the *appliance* dataset, the difference between the Interaction and Connection approaches is more subtle. The Interaction one is better in two cases; equal to the Connection in one and worse in another. For the *groceries megastore* dataset the Connection Graph outperforms the Interaction one in all situations.

This difference of performance of both approaches in the diverse datasets may be explained based on various facts. Firstly, for both *soda* and *appliance* datasets, the influential users were chosen by specialists in marketing and communication. The chosen users have a profile different from the ones chosen by the regular participants of the evaluation pool. Differently from the specialists, the participants of the pool were prone to classifying as influential users those with a higher number of followers (as briefly discussed in Section 4.2.2, while comparing the baseline results).

Having in mind that a high number of followers was a common attribute of influential users (especially evangelists), for the *groceries megastore* dataset, it makes

Table 4.7:  $\mathcal{F}_2^i$  values for the ranked lists. The arrows indicate the higher (best) ( $\blacktriangle$ ) and lower (worst) ( $\blacktriangledown$ ) values. The circle ( $\bullet$ ) indicates equal or approximated values. The parameters  $\alpha$ ,  $\beta$  and  $\gamma$  used in this experiment were determined in a *leave-one-out* procedure.

	$\mathcal{F}_2^e$	$\mathcal{F}_2^d$	$a(\mathcal{F}_2^e)$	$a(\mathcal{F}_2^d)$
<i>soda</i>				
$G_i$	1.00 $\blacktriangle$	0.05 $\blacktriangle$	1051.00 $\blacktriangle$	95.00 $\blacktriangle$
$G_c$	0.05 $\blacktriangledown$	0.04 $\blacktriangledown$	84.00 $\blacktriangledown$	85.00 $\blacktriangledown$
<i>appliance</i>				
$G_i$	0.52 $\blacktriangle$	0.11 $\bullet$	510.43 $\blacktriangle$	169.00 $\blacktriangledown$
$G_c$	0.07 $\blacktriangledown$	0.11 $\bullet$	136.00 $\blacktriangledown$	173.00 $\blacktriangle$
<i>groceries megastore</i>				
$G_i$	0.08 $\blacktriangledown$	0.23 $\blacktriangledown$	91.77 $\blacktriangledown$	317.42 $\blacktriangledown$
$G_c$	0.14 $\blacktriangle$	0.29 $\blacktriangle$	172.16 $\blacktriangle$	349.22 $\blacktriangle$

Table 4.8: Difference of recall ( $G_i - G_c$ ), with 90% confidence intervals. The symbol  $\blacktriangle$  highlights the cases in which  $G_i$  is better,  $\blacktriangledown$  highlights when  $G_c$  is better and  $\bullet$  shows the cases in which the difference between the approaches is not statistically significant.

	evangelists	detractors
<i>soda</i>	6.6667 $\pm$ 2.2187 $\blacktriangle$	13.3333 $\pm$ 2.9733 $\blacktriangle$
<i>appliance</i>	3.7681 $\pm$ 4.2180 $\bullet$	0.8333 $\pm$ 4.4112 $\bullet$
<i>groceries megastore</i>	-19.1667 $\pm$ 4.7390 $\blacktriangledown$	-3.3333 $\pm$ 1.8264 $\blacktriangledown$

sense for the  $\mathcal{F}_2^e$  values of the Connection Graph to be almost two times the values of the Interaction Graph. For the detractors, this difference is smaller, but the Connection Graph still wins.

### Paired Observations

Next, we provide a deeper comparison of the ranked lists generated using the Connection and Interaction Graph approaches. For this analysis, we employ a common procedure called *comparison of alternatives using paired observations* [Jain, 1991]. This procedure compares two or more systems in order to find the best among them.

The observations are called *paired* when, for two systems  $A$  and  $B$ , in the  $n$  experiments conducted, there is a one-to-one correspondence between the  $i$ -th test in system  $A$  and the  $i$ -th test in system  $B$ . The two samples, generated by the experiments on  $A$  and  $B$ , are treated as one sample of  $n$  pairs. The difference of performance is computed for each pair and a confidence interval is defined. The interval is used as means of checking if the difference measured is significantly different from zero, at a desired level of confidence. If it is, the systems are significantly different. The sign indicates which one has a better performance.

We apply this procedure for comparing both Graph approaches in the three datasets. We conducted 15 evaluations (*recall @  $x$* ,  $10 \leq x \leq 150$ ) consisting of paired observations of the experiments. The goal is to compare how many evangelists and detractors were retrieved using each approach, while the size of the ranked lists grows. We treat the samples of Interaction and Connection Graph as one single sample with 15 pairs and compute the difference for each one of them.

Figure 4.11 presents the values of evangelist’s and detractor’s *recall @  $x$*  for each approach and dataset. Table 4.8 presents the recall differences, with 90% confidence intervals. The Interaction Graph leads to better results in all cases for the soda dataset.

Firstly, for the soda dataset, the Interaction approach was significantly better. As mentioned before, many people mention soda brands on tweets in an occasional way.

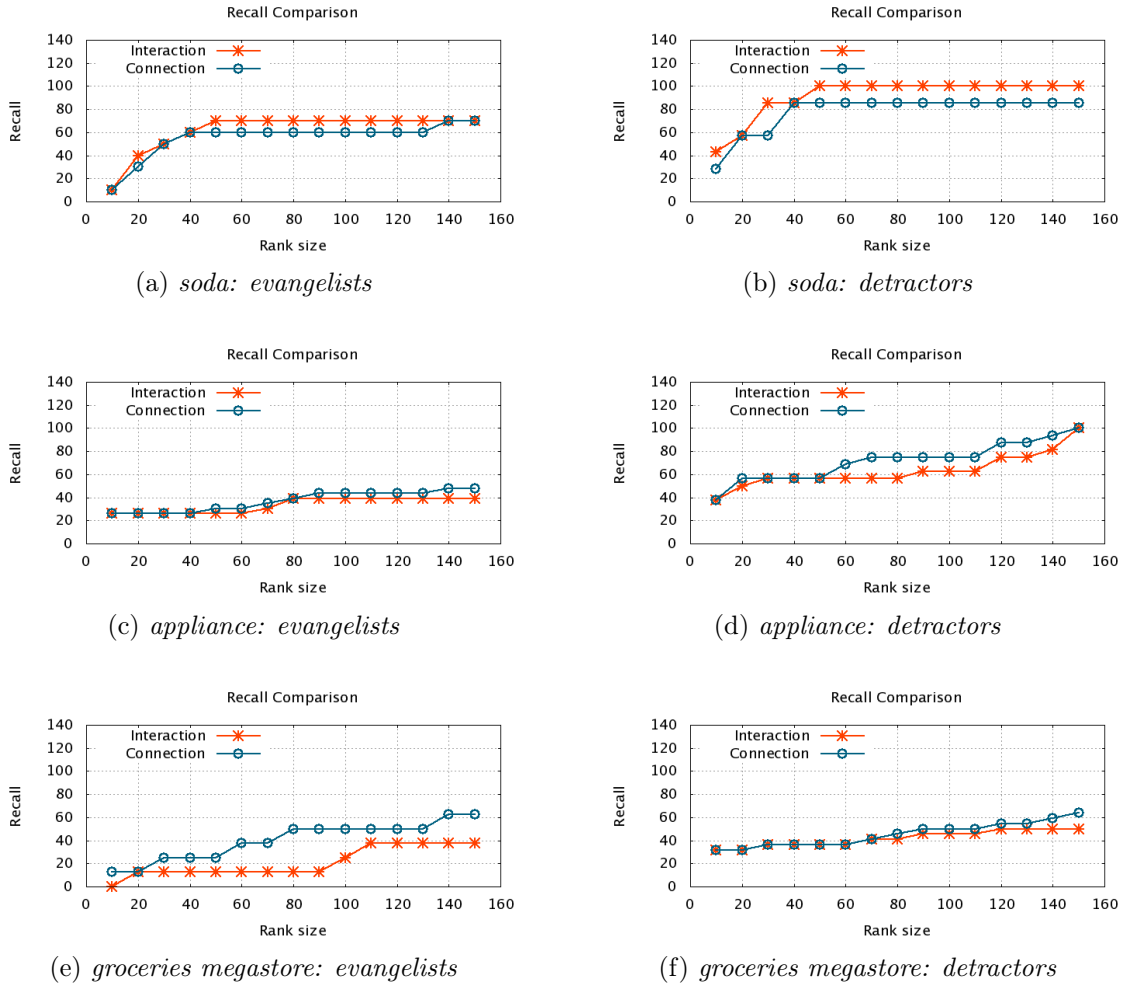


Figure 4.11: Paired observations for Interaction and Connection Graph approaches for evangelists and detractors' recall @  $x$  in both datasets. The parameters  $(\alpha, \beta, \gamma)$  of Formula 3.5 are calculated using a *leave-one-out* procedure.

Therefore, many users that may appear in the dataset are not truly trying to influence others, but their connections may mislead the results. For example, the connection network may indicate as influential a celebrity who cites a soda brand without the intention of influencing other people. Meanwhile, the Interaction approach is more precise at this point, and would only point out this celebrity as influential if her tweet had repercussion.

For the appliance dataset, the difference between the Interaction and Connection approaches is not statistically significant (the intervals include 0), which means that they lead to approximately the same result. We believe that both graph-based approaches have similar results in the appliance dataset due to its smaller size and the

non-usual characteristic of the topic. Since there are less users involved in the discussions about the brand, the chance that interaction will happen between two users that are connected is higher.

On the other hand, for the groceries megastore dataset, the Connection approach is better for both cases. The difference between the approaches is even higher on the evangelists dataset, probably due to the characteristics of the influential users identified in the pool, as discussed earlier.

### Computational Complexity

We also analyze the computational complexity of the extraction of *betweenness* (bc), *in-degree* (in) and *eigenvector centrality* (ec) for  $G_i$  and  $G_c$ , on each dataset<sup>4</sup>. We first sampled 100 results for each metric, in order to find the smallest number of observations  $n$  needed to provide an accuracy of  $\pm 10\%$ , with a confidence level of 90% [Jain, 1991].

Similarly to the Random Baseline equation, the number of observations required for each metric is given by Equation 4.8, in which  $\bar{x}$  is the mean of the observations,  $s$  is the standard deviation,  $r$  is the desired accuracy and  $z$  is the normal variate of the desired confidence level. That is,  $r = 10$  and  $z_{0.95} = 1.645$ . No metric needed a number of observations  $n$  higher than 100. Table 4.9 exhibits the average mean cost (in seconds) and the 90% confidence intervals for each case.

$$n = \left( \frac{100zs}{r\bar{x}} \right)^2 \quad (4.8)$$

For most of the cases, the difference of computing time between the two approaches was not statistically significant. Except for the eigenvector metric in both soda and appliance dataset, all the other differences' interval contained 0. That means that for only two cases the Connection Graph was worse than the Interaction one, in terms of computing time. So, we conclude that the computing time is not a relevant factor when choosing which approach to use.

Finally, an important and additional cost of  $G_c$  approach is to collect all the follower and following relations for the users in the dataset. Twitter API has limits of access, permitting up to 350 requests per hour, turning the pre-processing, in part, slow and expensive. For example, for the soda dataset, to get all the followers of each user (6885 users in total), no less than 19 hours would be necessary to collect the data. In light of the fact that many users have a large number of followers, more than one request per user is necessary and the cost for collecting the data is even higher.

---

<sup>4</sup>The system used for the experiments had the following configuration: Intel Core i7-2600 CPU @ 3.40GHz, 8GB RAM.

Table 4.9: Computing time comparison, in seconds, of betweenness and eigenvector centrality in  $G_i$  and  $G_c$ . The symbol  $\blacktriangle$  accounts for the cases in which  $G_c$  have a higher time of execution than  $G_i$  and  $\bullet$  for the cases in which the difference includes zero and both approaches have statistically equal time of execution.

	betweenness (bt)	indegree (in)	eigenvector (ec)
<i>soda</i>			
$G_i$	0.0111 (0.0020, 0.0202)	0.0065 (0, 0.0138)	0.1329 (0.0577, 0.2081)
$G_c$	0.0123 (0, 0.0179)	0.0072 (0, 0.0179)	3.8168 (3.3686, 4.2650)
$G_c - G_i$	0.0012 (-0.0166, 0.0190) $\bullet$	0.0006 (-0.0123, 0.0136) $\bullet$	3.6839 (3.2451, 4.1227) $\blacktriangle$
<i>appliance</i>			
$G_i$	0.0019 (0.0003, 0.0035)	0.0008 (0.0008, 0.0009)	0.0739 (0.0704, 0.0775)
$G_c$	0.0022 (0, 0.0045)	0.0010 (0.0005, 0.0015)	0.2415 (0.2290, 0.2539)
$G_c - G_i$	0.0003 (-0.0025, 0.0031) $\bullet$	0.0002 (-0.0003, 0.0006) $\bullet$	0.1676 (0.1546, 0.1805) $\blacktriangle$
<i>groceries megastore</i>			
$G_i$	0.0040 (0.0000, 0.0075)	0.0023 (0.0020, 0.0027)	2.0828 (1.9159, 2.2497)
$G_c$	0.0041 (0.0013, 0.0069)	0.0025 (0.0016, 0.0034)	1.9527 (1.7567, 2.1487)
$G_c - G_i$	0.0001 (-0.0026, 0.0027) $\bullet$	0.0002 (-0.0008, 0.0012) $\bullet$	-0.1301 (-0.3920, 0.1318) $\bullet$

Considering this cost, the size of the graphs and the approaches' performance when finding the influential users, we conclude that the Interaction Graph is better for determining user influence. It provides good (groceries megastore) or even better (soda and appliance) results than the Connection approach and it is simpler to construct.

## 4.4 Perspective Impact and Parameter Estimation

The second part of the experiments aims at discussing issues related to the parameters used in Equation 3.5, i.e.,  $\alpha$ ,  $\beta$  and  $\gamma$ . We analyze the impact of each perspective (relation, polarity and content) in the results of the method and propose a strategy to tuning those parameters.

### 4.4.1 The Impact of Each Perspective

In this first part of the parameters analysis, we evaluate the performance of each perspective in determining the influential users. Then, we study which are the perspectives responsible for the greatest fraction of variation of the influence detection results.

### An Overview of Each Perspective

As stated before, a single perspective (polarity, relation or content) may not be good enough to classify users as influential or not. In order to test this hypothesis, different rankings were generated using only one component of Formula 3.5 at a time. Figures 4.12, 4.13 and 4.14 present  $recall @ x$  results for each isolated component. We also present values for the baselines.

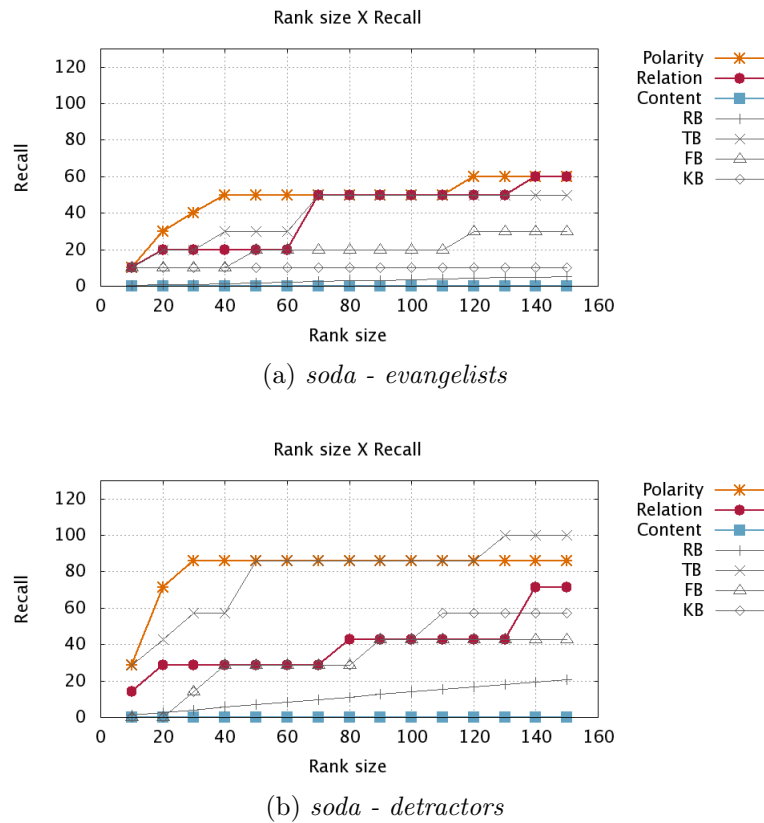
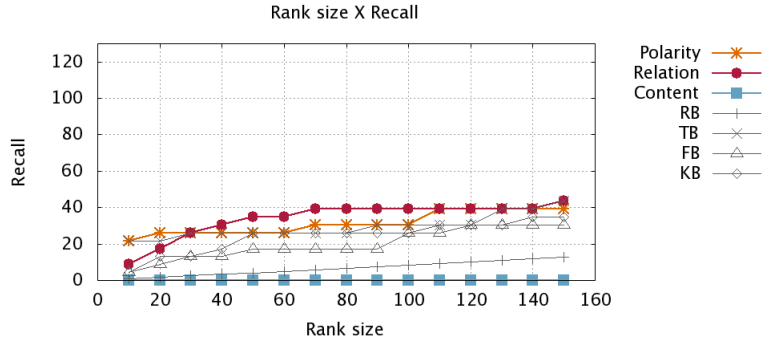
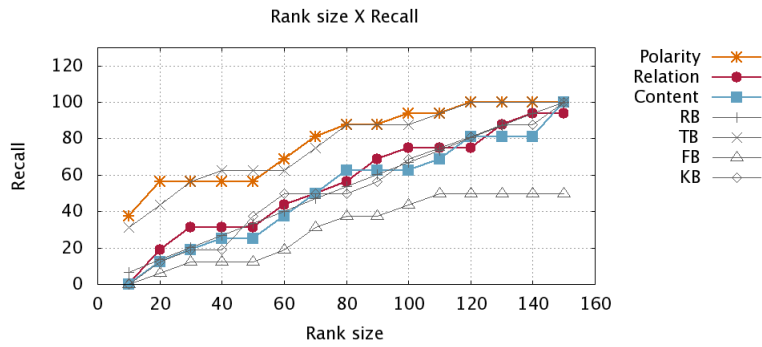


Figure 4.12: Plot of  $recall @ x$ , using  $G_i$ , considering only polarity, relation and content in both datasets. For *polarity* the parameters of Formula 3.5 are be  $\alpha = 1, \beta = \gamma = 0$ , for *relation*,  $\beta = 1, \alpha = \gamma = 0$  and for *content*,  $\gamma = 1, \alpha = \gamma = 0$ . Baseline curves are also displayed for each case, for comparison.

For the *soda* dataset, Figure 4.12, polarity by itself produces better results than the other two perspectives for both evangelists and detractors. Specifically, for detractors, polarity gives results significantly better than the other two perspectives whereas, for evangelists, the relation perspective is very close to polarity. Similarly, for the *appli* dataset, the relation perspective works better for evangelists (with a performance close to the polarity perspective), whereas polarity works better for detractors. In both datasets, besides the larger volume of positive tweets, analysts claim that the differ-



(a) *appliance - evangelists*



(b) *appliance - detractors*

Figure 4.13: Plot of  $recall @ x$ , using  $G_i$ , considering only polarity, relation and content in both datasets. For *polarity* the parameters of Formula 3.5 are  $\alpha = 1, \beta = \gamma = 0$ , for *relation*,  $\beta = 1, \alpha = \gamma = 0$  and for *polarity*,  $\gamma = 1, \alpha = \gamma = 0$ . Baseline curves are also displayed for each case, for comparison.

ence between neutral and positive tweets is quite subtle (which can lead to errors if one looks only at the polarity). For this reason, for evangelists detraction, the relation perspective outperforms polarity.

Still in the soda dataset, the relation perspective provides quite good results for both sentiments. It has a better performance than the majority of baselines (except TB on the detractors plot). The content perspective, on the other hand, does not provide any help on finding the influentials for the soda dataset. We believe that the low performance of the content perspective is probably due to the informal and noisy vocabulary used by Twitter users.

On the other hand, for detractors identification in the the appliance dataset, the content perspective by itself is practically as good as the relation perspective. As already mentioned, in the appliance dataset most of the negative tweets are from users who explore Twitter as *customer care* platform, reporting problems and dis-



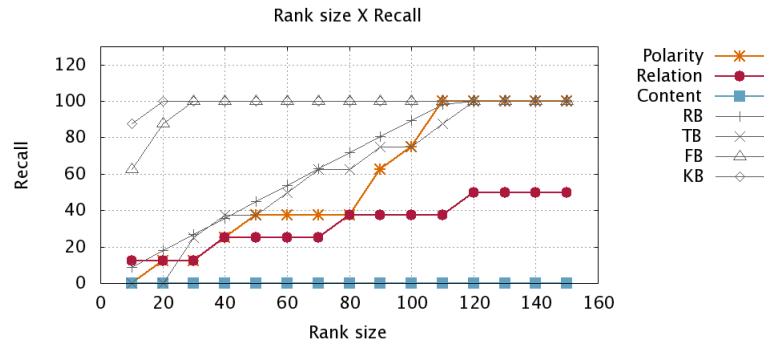
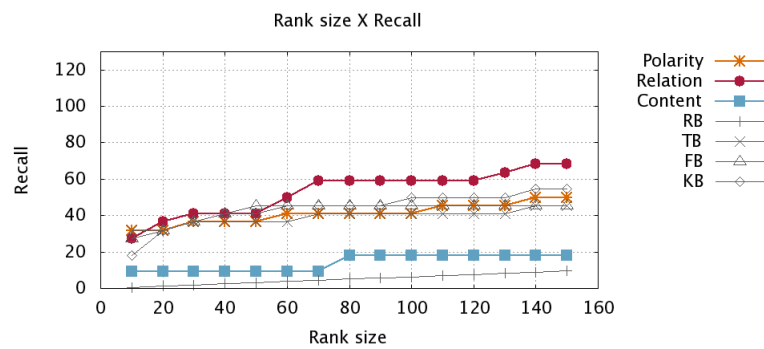
(a) *groceries megastore - evangelists*(b) *groceries megastore - detractors*

Figure 4.14: Plot of  $recall @ x$ , using  $G_i$ , considering only polarity, relation and content in both datasets. For *polarity* the parameters of Formula 3.5 are be  $\alpha = 1, \beta = \gamma = 0$ , for *relation*,  $\beta = 1, \alpha = \gamma = 0$  and for *polarity*,  $gamma = 1, \alpha = \gamma = 0$ . Baseline curves are also displayed for each case, for comparison.

satisfactions directly to the official brand profile. For such reason, we believe that the negative tweets are significantly well-written.

The groceries megastore dataset is quite different from the other two: there are very few representants of the positive class and the users classified as evangelists had the highest number of followers. Considering that, for this parameter analysis, the graph used as source for the relation perspective was the Interaction one, the relation perspective was not the best one this case. The polarity perspective had a fair result for the evangelists' detection, but reached 100% of recall only when the rank list size was greater than the number of positive users. Both TB and KB outperformed every perspective. They were the ones that best represented the follower relation of the user.

Finally, similarly to the others datasets, in the groceries megastore dataset, the relation feature had a better result for the sentiment with the highest number of instances: negative. When there is a large number of users in one sentiment, the relation

perspective is more efficient on discretizing influential and non influential users.

### Factorial Design

In order to perform a deeper analysis of the impact that each perspective has on the final method results, we employ a  $2^k$  *experimental (or factorial) design* [Jain, 1991].

In a experimental design, the outcome of an experiment is called the *response variable* and is the manner of measuring the system performance. Each variable that affects the response variable and has different alternatives is called a *factor* or *predictor* and its alternatives (the values it can assume) are called *levels*.

A *full factorial design* investigates every possible combination at all levels of all factors, determining the effect of  $k$  different factors (and inter-factor interactions) on the response variable. The number of factors and their levels can be very large and, consequently, the full factorial design may be expensive. Thus, there is a very popular design, called  $2^k$  design, in which each of the  $k$  factors is evaluated at two levels. This design acts as a preliminary investigation of which factors are relevant for a deeper investigation. The importance of a factor is measured by the proportion that it explains of the total variation of the response. In particular, the factors that explain a high percentage of variation are considered the most relevant for further investigation. The steps of an illustrative factorial design with two factors  $A$  and  $B$  can be summarized as follows.

#### $2^k$ Factorial Design Steps.

1. Each of their  $k$  factors are associated to variables  $x_A$  and  $x_B$ , that stands for the lower and higher levels, as follows:

$$x_k = \begin{cases} -1 & \text{if factor } k \text{ assumes its lower level,} \\ +1 & \text{if factor } k \text{ assumes its higher level.} \end{cases}$$

2. The performance (response variable)  $y$  of systems  $A$  and  $B$  are regressed on  $x_A$  and  $x_B$  using a nonlinear regression model of the form:  $y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B$ .
3. The effects  $q_0$ ,  $q_A$ ,  $q_B$  and  $q_{AB}$  are determined by expressions called *contrasts*, which are linear combinations of the responses  $y_i$  calculated based on observations of each possible combinations of the variables. If  $x_{Ai}$  and  $x_{Bi}$  are the levels of  $x_A$  and  $x_B$ , respectively, the observation would be modeled as:  $y_i = q_0 + q_A x_{Ai} + q_B x_{Bi} + q_{AB} x_{Ai} x_{Bi}$ .

4. The importance of a factor is measured by the proportion of the total variation in the response that is explained by the factor. In order to calculate this proportion, it is first, necessary to calculate the total variation of  $y$ , or the *sum of squares of total*, given by:  $SST = \sum_{i=1}^{2^k} (y_i - \bar{y})^2$ .
5. Also,  $SST$  can be expressed as  $SST = 2^k q_A^2 + 2^k q_B^2 + 2^k q_{AB}^2$ . The three parts on the right-hand side represent the portion of the total variation explained by the effect of  $A$ ,  $B$ , and interaction  $AB$ , such as  $SSA = 2^k q_A^2$ ,  $SSB = 2^k q_B^2$  and so on. Thus, the *fraction of variation explained* by a factor  $k$  is given by  $k = \frac{SSk}{SST}$ . Finally, this fraction provides means to gauge the importance of the factor.

For our experiment, we define the variables  $x_A$ , for polarity,  $x_B$ , for relation, and  $x_C$ , for content and the *response variable* is  $a(\text{recall})$ . The combination of factors was the following:

$$x_A = \begin{cases} -1 & \text{if } \alpha = \frac{1}{|u_{polarity}|} \\ +1 & \text{if } \alpha = 1. \end{cases} \quad x_B = \begin{cases} -1 & \text{if } \beta = 0 \\ +1 & \text{if } \beta = 1. \end{cases} \quad x_C = \begin{cases} -1 & \text{if } \gamma = 0 \\ +1 & \text{if } \gamma = 1. \end{cases}$$

For polarity, in the lowest level, only the signal of user's polarity is considered, while for the highest, the intensity is also taken into account. For example, considering a user with polarity perspective  $u_{polarity} = -0.12$ , in the lowest level (replacing  $\alpha$  in the influence score formula, Equation 3.5), the polarity part would be:

$$\begin{aligned} \alpha \times u_{polarity} &= \frac{1}{|u_{polarity}|} \times u_{polarity} \\ &= \frac{1}{0.12} \times -0.12 \\ &= -1. \end{aligned}$$

Meanwhile, in the highest level, the polarity part would be  $\alpha \times u_{polarity} = 1 * u_{polarity} = -12$ . For the relation and content perspectives, the levels were defined as the presence or absence of the component in the influence score formula ( $\beta = \{0, 1\}$  and  $\gamma = \{0, 1\}$ ). The intuition of employing this design is to analyze what is the effect on the results when a perspective can be left out.

In total, two scenarios were studied for each dataset, applying the described experimental design. In the first ( $D$ ), we considered the retrieval of detractors, in the second ( $E$ ), the retrieval of evangelists. Table 4.10 shows the results for each of the six designs by means of the *fraction of variation* for each factor for the datasets. The

perspective that turns out to be the most responsible for the variation either worsens or improves the results with much more intensity than the others.

By observing the results, we can conclude that the responsible for the greatest fraction of the *variation of results* for the first two datasets (soda and appliance) is the polarity perspective, represented by factor *A*. The use of the polarity signal, instead of its intensity, worsens the result largely. Also, as observed in the Figures 4.12 and 4.13, the polarity is one of the most important perspectives for these datasets.

For the soda dataset, content is the minor responsible for the variation for both evangelists and detractors. This means that the presence or absence of the metric does not impact the method much, that is, its contribution for influence detection is small.

Meanwhile, for the appliance dataset, content was responsible for a fraction of variation greater than the relation perspective for detractors, endorsing the aforementioned fact that the negative tweets on the appliance dataset are more well-written.

For the groceries megastore dataset, the results shown in Table 4.10 are interesting. For the evangelists, relation is responsible for  $\sim 76\%$  of the variation, followed by the interaction of relation and content, with  $\sim 24\%$ . As mentioned earlier, the sentiment of the tweets was not determinative when detecting evangelists on this dataset. Due to the participants’ bias towards users with higher number of followers, polarity did not impact as much in the influence detection. Meanwhile, network had a fairly important role. Since content by itself was not that important, we believe that the contribution of the *BC* factors was mainly due to the relation perspective. For detractors’ detection, though, polarity was responsible for a greater fraction of variation, along with relation.

Table 4.10: Factorial design results for both evangelist (E) and detractors (D) for both datasets.

Factorial Design Results								
<b>Soda</b>	Factors	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>
D	<i>% variation</i>	87.20%	5.60%	2.17%	2.21%	0.33%	2.14%	0.34%
E	<i>% variation</i>	41.26%	22.55%	7.02%	9.53%	9.86%	6.91%	2.87%
<b>Appliance</b>	Factors	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>
D	<i>% variation</i>	60.41%	9.05%	23.21%	0.04%	1.01%	6.29%	0.00%
E	<i>% variation</i>	49.91%	13.94%	13.28%	5.50%	8.83%	5.93%	2.62%
<b>Groceries megastore</b>	Factors	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>
D	<i>% variation</i>	25.54%	26.86%	2.61%	24.47%	2.50%	8.90%	9.11%
E	<i>% variation</i>	0.08%	76.03%	0.28%	0.04%	0.00%	23.56%	0.00%

### 4.4.2 Estimating the Parameters

Finally, determining the combination of  $\alpha$ ,  $\beta$ , and  $\gamma$  that provides the best result is an important issue. In this Section we analyze the combination of parameters and its effect on  $a(\text{recall})$  for detractors. Then we propose a way of tuning the parameters for the best results.

We start by presenting a ternary plot for the three parameters. Each edge corresponds to a parameter and its values increase vertically according to its opposite base, from 1 to 9, as Figure 4.15 illustrates. That is, on the top edge we have the combination  $\alpha = 1$ ,  $\beta = 1$  and  $\gamma = 9$ ; on the left edge,  $\alpha = 9$ ,  $\beta = 1$  and  $\gamma = 1$ ; and on the right edge,  $\alpha = 1$ ,  $\beta = 9$  and  $\gamma = 1$ . Moreover, in the middle of the triangle, we have  $\alpha = 5$ ,  $\beta = 5$  and  $\gamma = 5$ . Figure 4.16 shows plots for all datasets and graph approaches ( $G_c$  and  $G_i$ ). Each point of the ternary plots is a combination of the three of parameters  $\alpha$ ,  $\beta$  and  $\gamma$ . The color of each point indicates the area below the  $\text{recall}@x$  curve, that is,  $a(\text{recall})$ , for the combination of parameters that it represents. The scale goes from 0 to 150, according to a rainbow palette.

By analyzing the plots, one can see that the values of  $a(\text{recall})$  are different between the datasets. However, considering the datasets individually, the result does not change much for the possible combinations of  $\alpha$ ,  $\beta$  and  $\gamma$ .

The soda dataset values of  $a(\text{recall})$  are  $\sim 120$ , reaching  $\sim 130$  for lower values of  $\gamma$  and decreasing to  $\sim 80$  for lower values of  $\alpha$ . This behavior highlights the importance of the polarity factor in this dataset and the fact that content is not the best feature for this dataset. For the other two datasets, the result range is more homogeneous: for appliance, the values stay around  $\sim 100$  and for groceries megastore, around  $\sim 70$ .

We conclude that by choosing an intermediary combination of the parameters, one can guarantee good results for all datasets.

Finally, the high impact of the polarity factor shown in the factorial design experiment results (Section 4.4.1) is not evident in the ternary plots. This happens due to the differences in the  $\alpha$  parameter range evaluation: the factorial design experiments

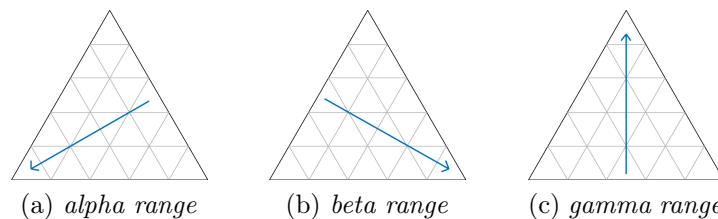
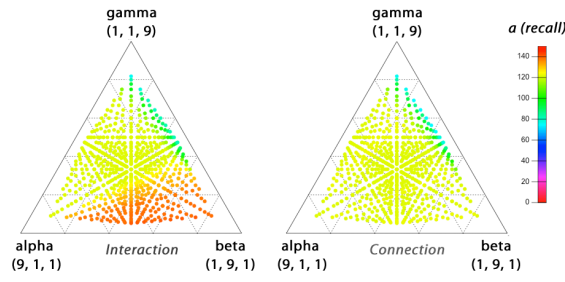
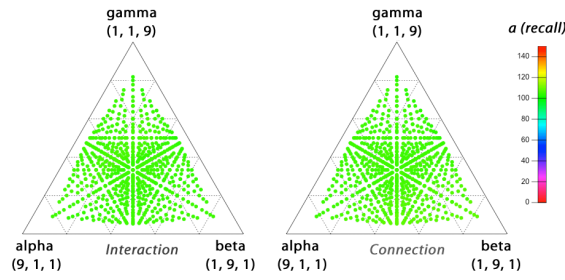


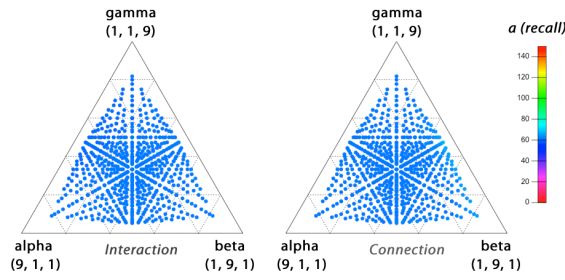
Figure 4.15: Plots showing the direction in which each parameter value changes.



(a) *soda dataset*



(b) *appliance dataset*



(c) *groceries megastore dataset*

Figure 4.16: Ternary plot of  $\alpha$ ,  $\beta$  and  $\gamma$  values for Interaction and Connection methods, for the different datasets. Each combination of parameters is a circle. The color (from a rainbow palette) represent the value for the area below the  $recall @ x$  curve:  $a(recall)$ , that goes from 0 to 150.

consider  $\alpha = \frac{1}{u|polarity|}$  and  $\alpha = 1$ , whereas the ternary plot analyzes  $\alpha$  with values from 1 to 9.

### Leave One Out

For estimating the potential of SaID, we optimize the parameters of the Influence Score formula. Specifically, we use a *leave one out* approach.

**Leave-one-out** is a type of cross-validation [Dietterich, 1998] in which a single observation is used as the test set, while the remaining observations are used as training data. Each observation is used only once as observation data. Actually, leave-one-out is a  $k$ -fold cross-validation with  $k$  being equal to the total number of observations in

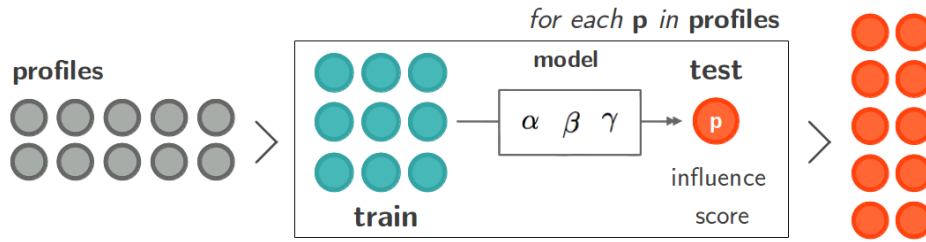


Figure 4.17: Leave one out for influence score.

the original sample.

In our case, we want to optimize the influence factor of each user, based on the parameters' choice. The following algorithm shows the steps of the *leave one out* procedure for finding the users' influence factor.

---

```

for  $p$  in  $profiles$  do
   $normalize(p)$ 
end for
for  $p$  in  $profiles$  do
   $train \leftarrow profiles - p$ 
   $(\alpha, \beta, \gamma) \leftarrow grid(train)$ 
   $I_{score}(p) \leftarrow test(p, (\alpha, \beta, \gamma))$ 
end for

```

---

Firstly, we normalize all the metrics so that the scores generated are comparable. Then, for each profile, we run `grid` for the remaining set of profiles (train). `grid` is a function that searches linearly all the parameters combinations from 1 to 9 and returns the  $(\alpha, \beta, \gamma)$  set that provides the best result in terms of  $a(recall)$  of detractors<sup>5</sup>. Using this parameter set, we calculate the influence score ( $I_{score}$ ) for the profile  $p$ . Figure 4.17 also illustrates the procedure.

## 4.5 Evangelists x Detractors

In this Section, we discuss the final results for ranking evangelists and detractors using the Interaction-based approach. We plot  $recall @ x$  curves for evangelists and detractors using SaID and all the baselines (Figure 4.18). We also run paired observations of SaID with each baseline, i. e. SaID x KB; SaID x TB; SaID x FB; and SaID x RB.

<sup>5</sup>We focus on detractors' detection with the aforementioned 'crisis management' approach in mind.

Table 4.11: Result of the paired observation of SaID with each baseline. We show the mean of the differences with their standard deviation and 90% confidence intervals.

evangelists				detractors				
<i>soda dataset</i>								
SaID - KB	42.0000	± 6.0027	13.2017	▲	60.9524	± 3.8559	8.4802	▲
SaID - TB	22.6667	± 4.3702	9.6115	▲	13.3333	± 3.8559	8.4802	▲
SaID - RB	52.0000	± 7.2291	15.8989	▲	56.1905	± 6.3845	14.0414	▲
SaID - FB	59.1193	± 7.7240	16.9874	▲	80.2632	± 5.7402	12.6245	▲
<i>appliance dataset</i>								
SaID - KB	13.3333	± 2.2990	5.0562	▲	32.0833	± 5.2474	11.5406	▲
SaID - TB	3.7681	± 2.2249	4.8933	▲	-13.7500	± 6.0950	13.4048	▼
SaID - RB	8.9855	± 1.6461	3.6203	▲	10.0000	± 8.1439	17.9109	▲
SaID - FB	26.6233	± 2.6382	5.8023	▲	9.2500	± 8.6463	19.0160	▲
<i>groceries megastore dataset</i>								
SaID - KB	-75.8333	± 5.4628	12.0144	▼	-0.6061	± 2.1910	4.8186	●
SaID - TB	-40.0000	± 11.2036	24.6403	▼	2.7273	± 1.5228	3.3490	▲
SaID - RB	-45.2480	± 5.4628	12.0144	▼	36.7282	± 2.3261	5.1157	▲
SaID - FB	-78.3333	± 10.1966	22.4254	▼	-2.1212	± 1.8043	3.9682	▼
<b>difference</b>	<b>mean</b>	<b>I.C.</b>	<b>stdev</b>		<b>mean</b>	<b>I.C.</b>	<b>stdev</b>	

We conducted 15 evaluations ( $recall @ x$ ,  $10 \leq x \leq 150$ ) for each pair and dataset, Based on the 15 evaluations, we computed the mean difference of performance in each scenario. In Table 4.11 we show the mean of the differences SaID - [baseline], the 90% confidence intervals and the standard variation. The interval is used as means of checking if the difference measured is significantly different from zero. If it is, the systems are significantly different. Positive values indicate that SaID had a better performance and negative values indicate that the baseline in comparison had a better performance.

SaID results are significantly better than every baseline for the soda dataset. Observing the plots for this dataset, the only case in which a baseline reaches the same recall that SaID does is the Tweet Baseline for detractors’ detection. Even then, while SaID reaches 100% of recall at the rank list size  $x = 50$ , the best baseline reaches the same recall only at  $x = 130$ . At the same time, no baseline reaches more than 50% of recall for the evangelists on this dataset.

For the appliance dataset, SaID is also significantly better than all the baselines, except for Tweet Baseline, also for detractors’ detection.



Table 4.12: Mean, standard deviation, minimum and maximum values for number of neutral tweets, positive tweets and followers for the evangelists.

	- tweets	+ tweets	followers
<i>soda</i>			
mean (stdev)	1 (1.70)	2.30 (1.06)	4009.71 (9591.91)
(min, max)	(0, 5)	(1, 4)	(102, 25753)
<i>appliance</i>			
mean (stdev)	0.17 (0.39)	3.65 (4.52)	3962.00 (8748.61)
(min, max)	(0, 1)	(1, 14)	(65, 34890)
<i>groceries megastore</i>			
mean (stdev)	0.25 (0.43)	1.00 (0.00)	8728.13 (9898.44)
(min, max)	(0, 1)	(1, 1)	(1094, 30130)

Meanwhile, for the groceries megastore dataset, on evangelists' detection, all the baselines outperformed SaID. As mentioned before, we believe that the choice of positive influential users was highly influenced by the number of followers of the users rather than their polarity or repercussion of content (features more explored by SaID). Moreover, due to the 'crisis' characteristic of the groceries dataset, we believe that the negative content and users may have overshadowed the positive content (there were only a few represents of the positive biased authors).

Table 4.12 shows the mean, standard deviation, minimum and maximum values for the number of positive tweets, neutral tweets and followers for the evangelists in the three datasets. In the groceries megastore collection, the number of followers is higher and the number of tweets lower than the other collections. The mean of followers for groceries megastore's evangelists is  $\sim 8700$ , with a high standard deviation of  $\sim 9900$  whereas for soda the mean is  $\sim 4000$  with standard deviation of 9600 and for appliance the mean is  $\sim 4960$  with standard deviation of  $\sim 8800$ . Examining the minimum and maximum values, while the minimum for soda and appliance are, respectively, 102 and 65 followers, for the groceries megastore is 1094. Meanwhile, the number of positive tweets is higher for the first two datasets than for groceries, endorsing our theory.

On the other hand, for detractors' detection, still in the groceries dataset, SaID performance was statistically equal to Klout's (the difference confidence interval include 0). Meanwhile, when compared to Tweet and Random Baselines, SaID was better. As expected, due to the ground truth characteristics, the Follower Baseline was better than SaID.

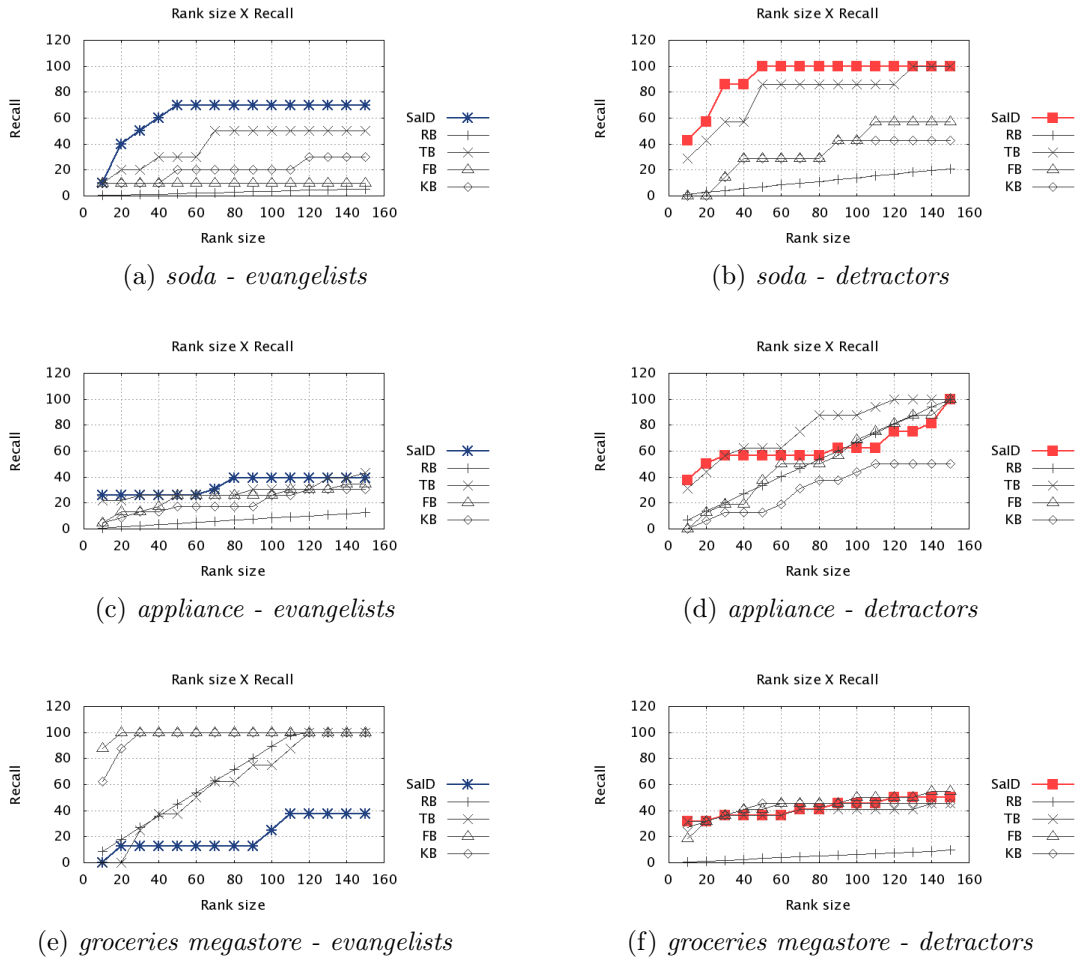


Figure 4.18: Comparison of SaID with the baselines for evangelists and detractors.

In summary, considering all the six combinations of datasets and polarities, SaID is the best method in three scenarios (soda-evangelist, soda-detractor, appliance-evangelist), it is the best or equal in two scenarios (appliance-detractor, groceries-detractor) and worst just in one scenario (groceries-evangelists). These results clearly demonstrate the effectiveness of our method for detecting influential users.

## 4.6 Concluding Remarks

This chapter addressed the experimental validation of SaID, our sentiment-based influence detection method.

We started by describing the collections built for evaluating our method. For each dataset, we discuss its statistics, the sentiment classification of its tweets and users and the ground truth of influential users. The three collections concern conversations about

brands: the first dataset is about soda, the second is about appliance and the third is about an groceries megastore chain. Even though all three datasets share the similar properties (they all consist on monitoring conversations concerning a product during an interval of time for marketing purposes) the datasets are very different from each other in terms of volume of content, number of active users, and type of posts. This dissimilarity enriches our analysis of influence.

We employed two methodologies for determining the ground-truth influential users used to evaluate our method. Specifically, for the two first datasets, soda and appliance, the influential users were identified by a group of specialists in marketing and communication, that provide this kind of service professionally. For the groceries megastore dataset, we determined the influential users based on a user study, in which non-specialist participants evaluated the profiles in a pool. This contrast of evaluation is directly reflected in our results. The type of user selected as influential by the specialists was different from the ones selected by the non-specialists. The later was very influenced by the number of followers of the profiles in the collections, which can be seen on the presented results. This difference of analysis may be related to the way the participants were instructed on what an influential user is. The instruction had a short explanation on influence which may be interpreted differently by each participant.

We also present, in this Chapter, the metrics and the baselines used to evaluate the influence detection results. We have implemented four baselines for comparing our method: one random, one based on Klout Score, one that orders user according to their number of tweets and one that orders by their number of followers. The intuition was to compare SaID with the most common ideas of influence. All the baselines are polarized, so that they all provide evangelists and detractors top users.

Our actual experiments were divided in three parts: (1) comparison of interaction and connection approaches; (2) analysis and estimation of the parameters of the Influence Score formula; and (3) evaluation of evangelists and detractors compared to the baselines. We have shown that SaID is efficient on detecting evangelists and detractors and that the interaction graph is the better choice for detecting influence.



## Chapter 5

# Experiments: Towards a Fully Automatic Approach

Lastly, we aim to evaluate the feasibility of a fully automatic approach, by examining the impact of an automatic sentiment analysis in the detection of the influential users.

As shown so far, SaID is a good method for detecting influential user on Twitter. However, a clear bottleneck of our method is the manual classification of the tweets' sentiment. The difficulty in automatically analyzing the sentiment of tweets have been addressed in many studies [Go et al., 2009, Pang et al., 2002, Read, 2005, Thelwall et al., 2010, Xia et al., 2011] and certainly these difficulties may impact our proposed influence detection method.

We used the manual classification of tweets for contrasting the graph approaches in Section 4.3, to perform the parameters analysis in Section 4.4 and for the evaluation of SaID performance compared to the baselines in Section 4.5. It is important now to take into account the impact of making the method fully automatic.

Accordingly, in this Section, we first present the results of the automatic classification of tweets and users. Then, we perform a detailed comparative analysis using paired observations of the best automatic result and the manually classified *skyline*, both using the Interaction graph.

### 5.1 Automatic Classification of Tweets

For analyzing the sentiment of the tweets, we used the state-of-art classifier SVM - Support Vector Machine [Joachims, 1998, Vapnik, 1995].

A **support vector machine** (SVM) is a binary supervised classifier. Given a training set in which each instance has a certain number of features and is assigned to one of two classes, SVM assigns one of those two classes to some instance that has similar features, but no known class.

Specifically, a SVM model is a representation of the training instances mapped as points onto a  $p$ -dimensional space (where  $p$  is the number of features), with a clear gap between the two categories. New instances are then mapped into this space, and their class is assigned based on which side of the gap the mapped point lies [Joachims, 1998, Vapnik, 1995]. We used the SVM implementation called LibSVM [Chang and Lin, 2001], and the parameters were defined using LibSVM's tool "Grid Parameter Search for Regression" on the train set. Specifically, for the three datasets, we used the parameters  $\gamma = 8$  and  $\text{cost} = 0.03125$ . We used the radial basis kernel.

We performed  $k$ -fold *cross-validation* [Dietterich, 1998], with  $k = 10$ . The 10-fold cross-validation technique divides the dataset  $\mathcal{X}$  into 10 equally (or approximately equally) sized parts or folds. Afterward, 10 iterations of training and validation are performed: in each iteration one fold is held-out as test while the remaining 9 are used for training, as illustrated by Figure 5.1.

Since the test folds do not contain intersection of instances, each tweet is classified only once. In such manner, to associate the classifier sentiment prediction with the tweets, we use the predictions for the test set, in each iteration.

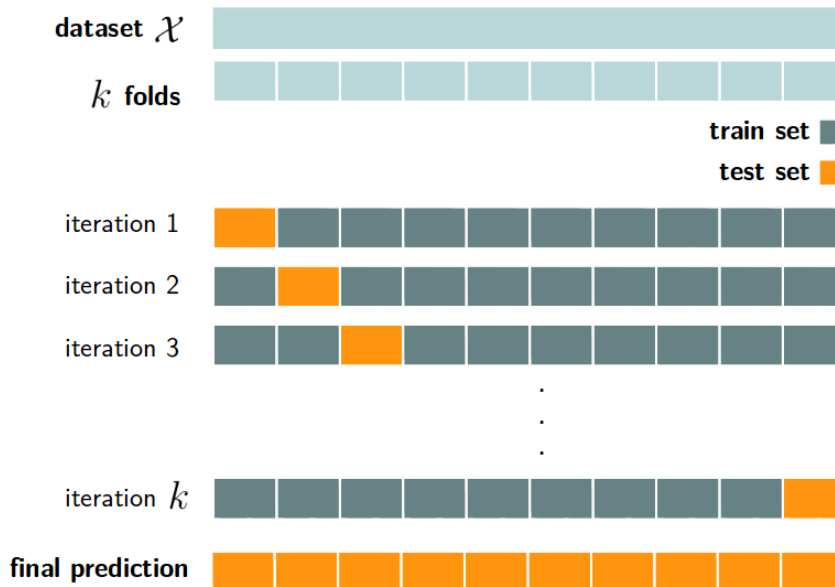


Figure 5.1: The 10-fold cross validation technique and final prediction of tweets.

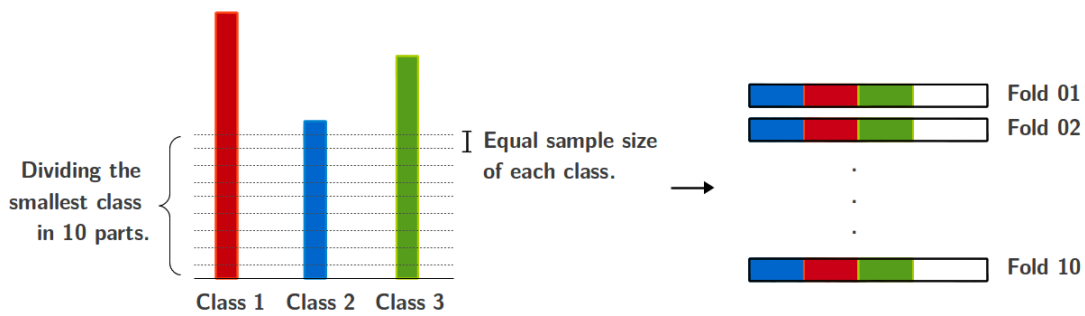


Figure 5.2: The 10-fold cross validation technique and final prediction of tweets.

As previously shown in Table 4.2, in the last Chapter, it is clear that there is a disparity on the number of instances of each class for all datasets. This imbalance of classes can lead to a decrease in the classification accuracy of the smallest class (negative tweets, in the soda dataset; positive tweet in the groceries megastore dataset) [Chawla, Nitesh V. et al., 2004]. This problem is known as the *class imbalance problem* and it is considered a challenge in the data-mining field [Yang and Wu, 2006]. Trying to workaroud this problem, the larger classes were undersampled, i. e., they had instances removed at random in order to reduce discrepancy between label counts [Prati and Monard, 2004]. Figure 5.2 illustrates the procedure. The smallest class instances are divided into 10 parts (10-fold cross validation). Each fold is filled with one of these parts. Then, for each of the other classes, we randomly sample the same number of instances for each fold, so that each fold has the same amount of instances of each class. Preliminary tests showed that classification results with this procedure were better than when using the original distribution.

We discuss the results of the SVM classifier using precision, recall,  $\mathcal{F}_1$  and Macro- $\mathcal{F}_1$ . Macro- $\mathcal{F}_1$  is the mean of  $\mathcal{F}_1$  for the different classes (positive, neutral and negative). For more details on evaluation metrics for classification, see Alpaydin [2004] and Mitchell [1997a]. All the values presented are the mean of the result in each fold and the intervals are calculated with 90% of confidence.

Observing Table 5.1, for the *soda dataset*, the recall values are very similar for the different types of sentiment, around 60%. On the other hand, the precision for the negative class is very low compared to the obtained for both neutral and positive classes. This means that the number of tweets classified as negative is way larger than the number of tweets that are really negative. As a consequence, the  $\mathcal{F}_1$  value is also very low for negative tweets. The value for Macro- $\mathcal{F}_1$  is  $0.55 \pm 0.01$ , which is low mainly due to the negative class.

As for the *appliance dataset*, the recall for negative tweets is better than the recall

Table 5.1: Tweet automatic sentiment classification results (with the 90% confidence interval).

	Positive	Neutral	Negative
<i>soda</i>			
<i>precision</i>	$0.64 \pm 0.02$	$0.71 \pm 0.01$	$0.27 \pm 0.01$
<i>recall</i>	$0.58 \pm 0.02$	$0.58 \pm 0.01$	$0.63 \pm 0.03$
$\mathcal{F}_1$	$0.61 \pm 0.02$	$0.63 \pm 0.03$	$0.38 \pm 0.02$
<i>appliance</i>			
<i>precision</i>	$0.86 \pm 0.01$	$0.46 \pm 0.02$	$0.44 \pm 0.04$
<i>recall</i>	$0.61 \pm 0.02$	$0.64 \pm 0.02$	$0.73 \pm 0.07$
$\mathcal{F}_1$	$0.71 \pm 0.02$	$0.53 \pm 0.01$	$0.55 \pm 0.04$
<i>groceries megastore</i>			
<i>precision</i>	$0.13 \pm 0.04$	$0.81 \pm 0.05$	$0.93 \pm 0.02$
<i>recall</i>	$0.71 \pm 0.28$	$0.77 \pm 0.04$	$0.86 \pm 0.03$
$\mathcal{F}_1$	$0.21 \pm 0.07$	$0.79 \pm 0.03$	$0.89 \pm 0.02$

for neutral and positive tweets. The precision, however, is lower for both negative and neutral tweets and higher for positive ones (that represent 63.25% of the tweets. The value for Macro- $\mathcal{F}_1$  is  $0.59 \pm 0.02$ , with, which is better than the *soda dataset* result.

Finally, for the *groceries megastore dataset*, the lowest *precision*, *recall* and  $\mathcal{F}_1$  values are for the positive class. This category represents only 1.41% of the tweets and the most part of the positive tweets were classified as neutral. The value for Macro- $\mathcal{F}_1$  is  $0.63 \pm 0.03$ . It is one of the highest Macro- $\mathcal{F}_1$  of the three datasets, mainly due to the high values for both negative and neutral classes.

## 5.2 From Tweets to Users

For detecting evangelists and detractors, the final polarity assigned to the user is more important than the accuracy of tweets classification. There is no damage in influential users detection if the overall polarity of each user is maintained. For example, the classification algorithm may predict a positive tweet as neutral and even then, the polarity of the user remain positive.

As a first effort to analyze the impact of the automatic classification on SaID, we present in Table 5.2 the *confusion matrix* of user polarity attribution using the



Table 5.2: Confusion matrix of the user polarity attribution. Each column represents the users whose polarity was calculated based on the automatic classification. Each row represents the instances whose polarity was calculated using the manual classification.

	soda			appliance			groceries megastore				
	+	o	-	+	o	-	+	o	-		
+	<b>1657</b>	843	96	+	810	<b>91</b>	22	+	79	396	<b>86</b>
o	<b>670</b>	1904	176	o	254	<b>215</b>	34	o	29	2124	<b>258</b>
-	<b>443</b>	654	442	-	134	<b>54</b>	93	-	4	158	<b>1238</b>

Table 5.3: Values of *precision*, *recall*,  $\mathcal{F}_1$  and Macro- $\mathcal{F}_1$  for profile attribution using automatic classification.

	soda			appliance			groceries megastore				
	<i>prec</i>	<i>recall</i>	$\mathcal{F}_1$	<i>prec</i>	<i>recall</i>	$\mathcal{F}_1$	<i>prec</i>	<i>recall</i>	$\mathcal{F}_1$		
+	0.64	0.60	0.62	+	0.88	0.68	0.76	+	0.14	0.71	0.23
o	0.69	0.56	0.62	o	0.43	0.60	0.50	o	0.88	0.79	0.83
-	0.29	0.62	0.39	-	0.33	0.62	0.43	-	0.88	0.78	0.83
	Macro- $\mathcal{F}_1 = 0.54$			Macro- $\mathcal{F}_1 = 0.56$			Macro- $\mathcal{F}_1 = 0.63$				

automatically analyzed tweets. A **confusion matrix** is a table layout that allows the visualization of the performance of an algorithm. Each column represents the instances in a predicted class (polarity based on the automatic classification) and each row represents the instances in an actual class (polarity based on the manual classification). This visualization is useful to identify mislabeling of classes.

Table 5.3 shows *precision*, *recall* and  $\mathcal{F}_1$  values for the polarity attribution using the automatic classification of tweets. *Recall* for a class  $c$ , calculated based on a confusion matrix, is the proportion of  $c$  cases that were correctly identified. *Precision*, on the other hand, is the proportion of the predicted  $c$  cases that were correct.

The recall values are homogeneous from an intra-dataset point-of-view. That is, recall for positive, neutral and negative users are close to each other, for each dataset. This means that for the soda and appliance dataset at least  $\sim 60\%$  of the users for each class were correctly classified, with the automatic analysis of tweets. For the groceries megastore dataset this proportion is about  $\sim 70\%$ . However, the precision values for the classes with fewer representants (negative users for soda and appliance dataset; positive users for groceries megastore dataset) were really low. This means that a great number of users was erroneously assigned as negative in soda and appliance

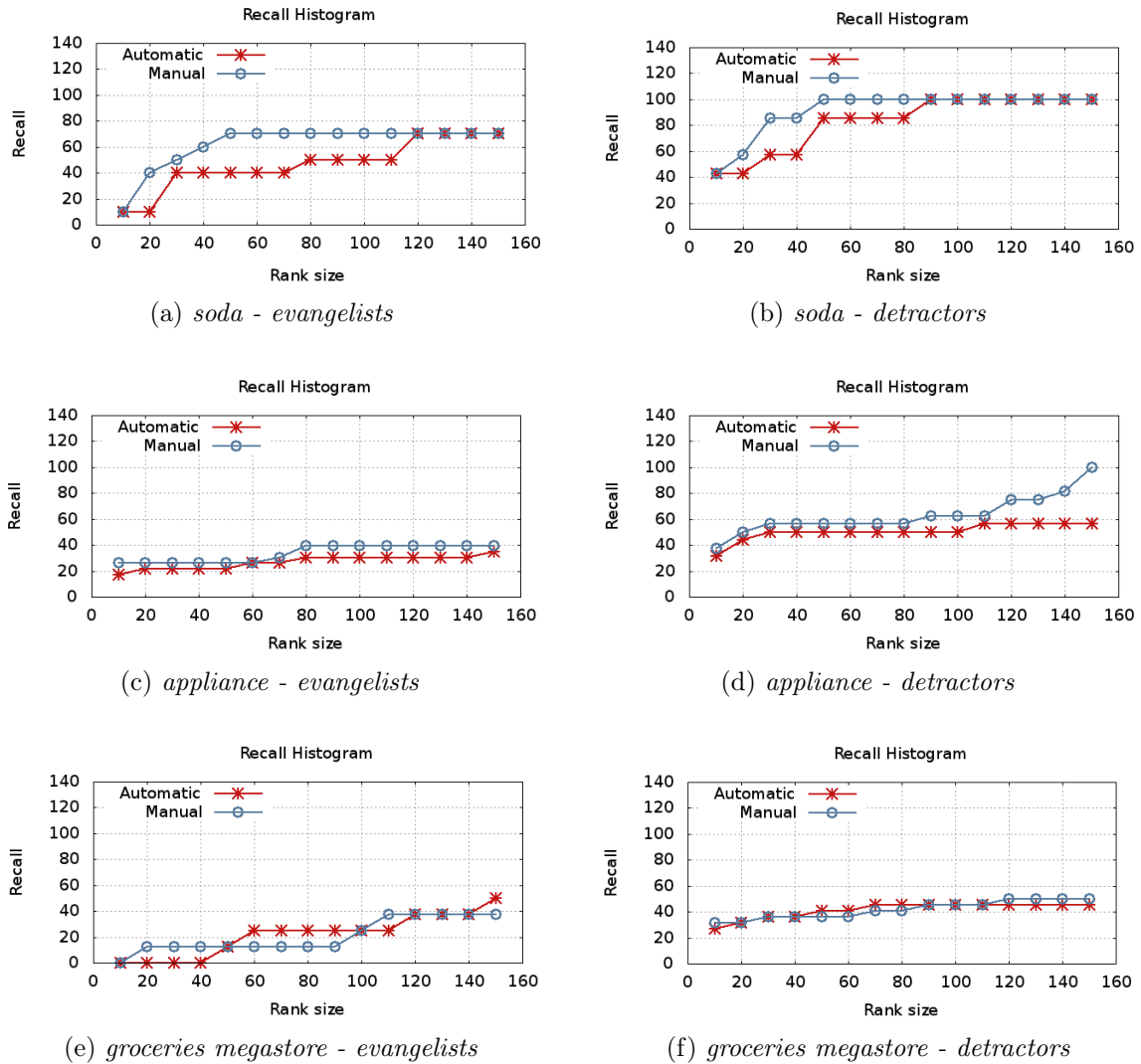


Figure 5.3: Comparing automatic and manual approaches.

datasets or erroneously assigned as positive in the groceries megastore dataset. This reflects on the  $\mathcal{F}_1$  value of these smallest classes and on Macro- $\mathcal{F}_1$  for each dataset.

### 5.3 Manual $\times$ Automatic Classification

Finally, we compare SaID results using both manual and automatic analysis of tweets. For comparing these two approaches, we conduct *paired observations*. Similarly to the other experiments, we conducted 15 evaluations of  $recall @ x$ ,  $10 \leq x \leq 150$  consisting of paired observations of the experiments.

The goal is to compare how many evangelists and detractors were retrieved using

either the manual or automatic analysis. We treat the samples of manual and automatic classification as one single sample with 15 pairs and compute the difference for each one of them.

Figure 5.3 presents the values of evangelists and detractors  $recall @ x$ . Analyzing the results, one can clearly see that the manual and automatic approaches are really similar to each other, but for some sizes of ranks the manual approach is slightly better.

The overall difference of the two methods for detractors, considering from TOP-10 to TOP-150 is not high. For the soda dataset, the mean recall difference of detractors is  $8.57 \pm 4.79$  (which is approximately the recall for detecting **one** detractor). For evangelists, the mean is  $15.33 \pm 5.67$  (which is approximately the recall for detecting **one or two** evangelists). After rank size 90 for detractors and 120 for evangelists, both approaches have equal results.

On the other hand, for the appliance dataset, the mean detractor's recall difference for the presented TOP sizes is  $12.50 \pm 4.80$  (which is approximately the recall for detecting **two** detractors) For appliance dataset's evangelists, the mean difference is only  $6.38 \pm 1.27$  (between **one and two** evangelists).

Finally, for the groceries megastore dataset, for detractors, the mean is  $-0.30 \pm 1.65$ , that is, both approaches are statistically equal. Similarly, for evangelists, the mean is  $0.83 \pm 4.54$ . The intervals presented have 90% of confidence.

## 5.4 Concluding Remarks

In this Chapter, we addressed the problem of automatizing the classification of the tweets' sentiment. Specifically, we evaluated the impact of this automatic analysis on our influence detection method.

Observing the results, we note that the difference of effectiveness between the automatic and the manual approach is rather small. Although there is an impact on SaID's result, it is not that high, showing that the method can be fully automatized without significant effectiveness loss.



# Chapter 6

## Conclusion

This dissertation addressed the problem of identifying biased influential users about a topic on Twitter. Motivated by the dynamics of this environment, in which users share opinions, experiences and suggestions about diverse subjects, and by the huge volume of content generated daily, we aim to assist businesses (or anyone interested in products or services feedback) on finding the key users who lead the conversations and actions for a given subject. Specifically, we list potential evangelists and detractors for a topic of interest. Our method, called SaID (Sentiment-based Influence Detection on Twitter) focuses on users' behavior when classifying them as influential or not. We extract features from the users, such as the polarity and readability of their tweets and their centrality in terms of interactions via tweets or following connections.

Concluding this dissertation, we list, in this final Chapter, our main contributions and what we plan to implement as future work.

### 6.1 Contributions

The main contributions of this dissertation were:

- **A new definition of “influential user” on Twitter.** In Section 3.1, we refined the concept of influential user, or opinion leader. The profiles who fit into this category are the ones responsible for producing an effect on other users. Particularly, their actions imply in other persons' actions; they act like bridges on the interactions about the topic; they have the intention of persuading the others positive or negatively; and they also produce content with a minimum expected quality. Besides defining what an influential user is, we explained the intuition

behind each characteristic and described how one can measure it using the data available on Twitter.

- **SaID, a method for detecting influential users on Twitter.** We presented in this dissertation a method called SaID - Sentiment-based Influence Detection on Twitter (Section 3.4). The method combines different user features into an Influence Score, that categorizes the user as an evangelists or detractor. We divide the features into three perspectives: polarity, relation and content. Polarity measures the sentiment bias on an users' tweets; relation captures users' position among the interactions via tweets or following connections; and, finally, content addresses the readability of users' tweets, trying to identify well written content. The intuition is that a user that is central in terms of interaction or connections about the topic, has a biased opinion and writes high content tweets should be ranked higher as an influential user.
- **Datasets fully analyzed that may act as benchmark for influence detection.** There is no established benchmark for influence detection evaluation on Twitter. Therefore, a great effort of this work was to build such collections. We have constructed three datasets, fully analyzed in terms of tweet sentiment and user influence (Section 4.1).
- **Detailed comparison of the effectiveness of interactions via tweets and following connections on determining influence.** We took into consideration two types of relations between users: the explicit ones, that happen through follower-following relations; and the implicit ones, that occur through interactions (mentions, replies, retweets) via posts. We have deeply analyzed the impact of each of these types of relations in user influence detection and found that the implicit interactions via tweets provide the best mechanism to determine influence (Section 4.3).
- **A thorough discussion on how polarity, relation and content may affect influential detection.** Our definition of influence conjectured that a user must have high polarity, relation and content features in order to be influential. In Section 4.4, we studied the performance of each perspective separately, for the task of finding influential users. Also, we carefully studied the feature set of the users in order to find the perspectives responsible for the highest impact on SaID results.

- **Considerations about the effect of automatic tweet classification on influence detection.** When trying to detect evangelists and detractors for a topic, a high quality of the sentiment classification of tweets is important, but even with errors on the classification, the overall polarity of the user may not be affected. Section 5 discussed the effects of automatic classification on sentiment-based influence detection.

In addition to the contributions listed above, this dissertation resulted in 2 papers: *Detecting Evangelists and Detractors on Twitter* [Bigonha et al., 2010], best paper in WebMedia 2010 and *Sentiment-based Influence Detection on Twitter* [Bigonha et al., 2011] in the Journal of Brazilian Computer Society.

## 6.2 Future Work

Following, we present a few issues left for future work:

- **Test machine learning as an alternative to rank the users.** There are some limitations using a formula for combining the different metrics of an user. By using an automatic approach for ranking users, we avoid making decisions such as how to combine the features and which metric should have a higher weight. Preliminary tests using the implementation of Naive Bayes [Mitchell, 1997b] of Weka Machine Learning Project [2009] showed good results for classifying the influentials based on users' features.
- **Test rank aggregation as an alternative to rank the users.** The Rank Aggregation Problem [Dwork et al., 2001] is to combine different rank orderings on the same set of alternatives in order to obtain a "better" ordering. We plan to test this method for combining the different ranks generated by the diverse sets of features into one influence rank.
- **Implement a better content perspective.** Not only readability indicates the good or bad textual quality of tweets. Castillo et al. [2011], for instance, showed ways to determine the credibility of the information available on Twitter. This aspect may be useful for determining influential users, because they tend to be trustworthy.
- **Take into account temporal aspects when determining influence.** One of the main characteristics of Twitter is the high speed with which information

changes. Although becoming opinion leaders or influential users, Twitter's immediatist property may accelerate this process. Thus, as future work, we plan to improve SaID so it can adapt the influential rank lists according to time.

- **Improve the tweet sentiment classification method.** As shown in Section 5, the erroneous classification of the sentiment of tweets may affect SaID results. As future work, we address the improvement of our sentiment classification strategy, for example, using ensemble methods [Xia et al., 2011].
- **Automatic content filtering.** Implement an automatic approach for filtering the inappropriate content retrieved by the crawler, in order to have a completely automatic method.



# Appendix A

## Characterizing the Content

In this Appendix, we intend to qualify the content of the datasets explored in this dissertation, for a better understanding of our discussions.

We first present the term clouds for each dataset and sentiment (Section A.1). Then, we list examples of tweets of the ground truth evangelists and detractors for each dataset (Section A.2). Since the datasets contain tweets written in Brazilian Portuguese, all the term clouds and tweets examples are in this language.

### A.1 Term Cloud

A term cloud is a visualization of word frequency of a given text. The higher the frequency of the word in the text, the biggest is its font size. In order to characterize the content for each combination of dataset and sentiment, we display the term cloud of all the tweets that characterize the corresponding set on Figures A.1, A.2 and A.3. The keywords used for crawling each dataset (the name of the brands) were removed from the term cloud for a better visualization.

For the soda dataset (Figure A.1), the positive content term cloud has terms like "better", "cold", "I want", "drinking", etc. The negative term cloud has terms like "not", "pepsi", "better", "looses", "gastritis", "bad" and other words giving the idea of comparison of the soda brand with others and analysis of the health impact of soda in general. Finally, the neutral term cloud has words like "drinking", "now", "day", "home", "eat", which give the idea of people just mentioning the soda brand as a part of their daily activities.

Next, for the appliance dataset, both positive and neutral term clouds contain words about a marketing campaign of the appliance brand: "campaign", "comercial" and "inspiration changes everything" (the name of the campaign). Particularly, the

positive term cloud also has terms such as "congratulations" and "good", whereas the neutral one has several URL links for the campaign video or repercussion (an informative approach). On the other hand, on the negative term cloud, the consumer service conversations are evident by the most frequent words: "problem", "fridge", "assistance" and "nothing". Also, "consul", an important competitor of the studied brand, appears with high frequency.

Finally, for the groceries dataset, the positive term cloud most frequent words are "bought", "sale", "tip", and so on. Meanwhile, the neutral term cloud has terms like "today", "here", "buy", "going", "at", which indicate that the users are mentioning the groceries megastore chain in their daily activities (coming or going to a store, buying something, etc). At last, the negative term cloud clearly indicates the problem with the purchases on the groceries online store. The most frequent words are: "order", "mega saldão" (the name of the sale event), "canceled", and so on.

## A.2 Datasets' Example Tweets

In order to illustrate our discussion about the influential users and their tweets, we now list some of the positive and neutral tweets of evangelists and detractors for each dataset.

### Soda Dataset

#### Evangelists' Tweets

- "Pra quem n acompanhou o estouro da Coca Cola com o abridor improvavel (despertador) veja o video aqui: <http://migre.me/649H>. Sim, so colono!"
- "@eclribeiro Pior você, que depois de uma árdua pedalada, não pode nem matar a sua sede com uma deliciosa coca-cola."

#### Detractors' Tweets

- "Saiba oque acontece com o seu organismo depois de tomar uma Coca-Cola - <http://tinyurl.com/nqkyfz> #interessante"
- "amanhã vou entrar em contato com a Coca Cola e denunciar fraude na promoção deles, estou com uma tampinha já cadastrada no site. #cocacola"

## Appliance Dataset

### Evangelists' Tweets

- "Hj foi dia de lançamentos Brastemp e Consul! Produtos fantásticos, legal ver a alegria da galera responsável por esse monte de coisas!..."
- "Assistam amanhã no intervalo do Jornal Nacional a nova campanha da Brastemp... Emocionante..."

### Detractors' Tweets

- "@brastemp Boas campanhas, péssima qualidade e atendimento...assim é a #brastemp"
- "A @brastemp tem um péssimo atendimento ao consumidor. Estou em contato há mais de 15 dias e nada de instalar minha coifa! #brastemp #fail"

## Groceries Dataset

### Evangelists' Tweets

- "Desconto na Churrasqueira Elétrica Mister Grill Plus Cotherm no Carrefour. De: R\$ 79,00 Por: R\$ 49,90. <http://t.co/sxEAsnNJ>"
- "Tv de 50polegadas na promoção imagina se não ficamos foda kkk vlw carrefour!!!"

### Detractors' Tweets

- "Mega Saldão do Carrefour se mostrou uma enganação. Cancelaram meus pedidos sem uma explicação razoável. @carrefourcombr @carrefourcombr"
- "@carrefourcombr Não comprem nada desse site. Estão cancelando as compras e desrespeitando todos os consumidores do MegaSaldao"





(a) *groceries - positive*(b) *groceries - neutral*(c) *groceries - negative*

Figure A.3: Term cloud for positive, negative and neutral tweets for the groceries dataset.

# Bibliography

- Ethem Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004. ISBN 0262012111.
- Johan Arndt. Role of Product-Related conversations in the diffusion of a new product. *Journal of Marketing Research*, 4(3):291–295, 1967. ISSN 00222437.
- Michael Arrington. Odeo Releases Twtr. *TechCrunch*. <http://techcrunch.com/2006/07/15/is-twtr-interesting/>, July 2006. URL <http://techcrunch.com/2006/07/15/is-twtr-interesting/>.
- Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 020139829X.
- Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone’s an Influencer: quantifying influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 65–74, Hong Kong, China, 2011. ACM Press. ISBN 9781450304931.
- Albert-László Barabasi. *Linked: the new science of networks*. Basic Books, 2002. URL [http://scholar.google.com/scholar?q=related:SIWWRt-RhYJ:scholar.google.com/&hl=en&num=30&as\\_sdt=0,5](http://scholar.google.com/scholar?q=related:SIWWRt-RhYJ:scholar.google.com/&hl=en&num=30&as_sdt=0,5).
- John A. Barnes. Classes and Committees in a Norwegian Island Parish. *Human Relations*, 7:39–58, 1954.
- Jonathan Berry and Edward Keller. *The Influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster, 2003. URL [http://books.google.com/books?hl=en&lr=&id=sI50vvhwdIOC&oi=fnd&pg=PA1&dq=the+influentials+one+american+in+ten+tells+the+other+nine+how+to+vote+where+to+eat+and+what+to+buy&ots=uJUo\\_23eY5&sig=jTWv4M3xVaGsEAIxT1YDksuw0qI](http://books.google.com/books?hl=en&lr=&id=sI50vvhwdIOC&oi=fnd&pg=PA1&dq=the+influentials+one+american+in+ten+tells+the+other+nine+how+to+vote+where+to+eat+and+what+to+buy&ots=uJUo_23eY5&sig=jTWv4M3xVaGsEAIxT1YDksuw0qI).

- Carolina Bigonha, Thiago N. C. Cardoso, Mirella M. Moro, Virgilio Almeida, and Marcos A. Gonçalves. Detecting evangelists and detractors on twitter. In *Simpósio Brasileiro de Sistemas Multimídia e Web - Webmedia 2010*, pages 107–114, Belo Horizonte, Minas Gerais, Brazil, 2010.
- Carolina Bigonha, Thiago N. C. Cardoso, Mirella M. Moro, Virgilio Almeida, and Marcos A. Gonçalves. Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, 2011. doi: 10.1007/s13173-011-0051-5.
- Phillip Bonacich. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555 – 564, 2007. ISSN 0378-8733. URL <http://www.sciencedirect.com/science/article/B6VD1-4NN6TDV-1/2/e56b352509c2c3030f232a8d5f2f3889>.
- Stefan Bornholdt and Heinz Georg Schuster. Handbook of graphs and networks. Wiley-VCH, 2003. URL <http://onlinelibrary.wiley.com/doi/10.1002/3527602755.fmatter/summary>.
- Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30, 2008. URL <http://www.sciencedirect.com/science/article/pii/S0378873307000731>.
- Alex Braunstein. Why your Klout score is meaningless, June 2011. URL <http://alexbraunstein.com/2011/06/01/why-your-klout-score-is-meaningless/>.
- Serge Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 1998. URL <http://www.citeulike.org/group/1374/article/802616>.
- Jacqueline J. Brown and Peter H. Reingen. Social Ties and Word-of-Mouth Referral Behavior. *The Journal of Consumer Research*, 14(3):350–362, 1987. URL <http://dx.doi.org/10.2307/2489496>.
- Jo Brown, Amanda J Broderick, and Nick Lee. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 21(3):2–20, January 2007. URL <http://linkinghub.elsevier.com/retrieve/pii/S1094996807700300>.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 2011. URL <http://portal.acm.org/citation.cfm?id=1963500>.



- Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, pages 1–10, New York, New York, USA, 2010. ACM Press. ISBN 9781450302203.
- Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, District of Columbia, USA, 2010. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1538/1826>.
- Kenny K. Chan and Shekhar Misra. Characteristics of the opinion leader: A new dimension. *Journal of Advertising*, 1990. URL <http://www.jstor.org/stable/4188770>.
- Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Chawla, Nitesh V. et al. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6:1–6, June 2004. ISSN 1931-0145.
- Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and Tweet: Experiments on Recommending Content from Information Streams. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, pages 1185–1194, New York, New York, USA, 2010. ACM Press. ISBN 9781605589299. URL <http://portal.acm.org/citation.cfm?doid=1753326.1753503>.
- P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google’s PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, January 2007.
- Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006. URL <http://www.atypon-link.com/AMA/doi/abs/10.1509/jmkr.43.3.345>.
- Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on Twitter: human, bot, or cyborg? In *Proceedings of the 26th Computer Security Applications Conference*, pages 21–30, New York, New York, USA, 2010. ACM Press. ISBN 9781450301336. URL <http://portal.acm.org/citation.cfm?doid=1920261.1920265>.

- Luciano F. Costa et. al. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007.
- Daniel Hasan Dalip et.al. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL'09*, pages 295–304, 2009. ISBN 978-1-60558-322-8.
- Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark My Words! Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web*, pages 141–150, May 2011. URL <http://arxiv.org/abs/1105.0673v1>.
- Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international Conference on Human Factors in Computing Systems*. ACM Request Permissions, April 2010. URL <http://portal.acm.org/citation.cfm?id=1753504>.
- Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10:1895–1923, 1998.
- Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 613–622, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.372165. URL <http://doi.acm.org/10.1145/371920.372165>.
- James Engel, Roger D Blackwell, and Robert J Kegerreis. How Information is Used to Adopt an Innovation. *Journal of Advertising Research*, 9:3–8, 1969.
- Nicholas Evangelopoulos and Lucian Visinescu. Text-Mining the voice of the people. *Communications of the ACM*, 55(2):62–69, February 2012. URL <http://cacm.acm.org/magazines/2012/2/145399-text-mining-the-voice-of-the-people/fulltext#.TygN09WwZfA.citeulike>.
- J L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. URL <http://www.psych.umn.edu/faculty/waller/classes/meas08/Readings/Fleiss1971.pdf>.
- Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, January 2002. ISBN 0316346624.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, 2009. URL <http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf>.

Jennifer Golbeck and Derek Hansen. Computing Political Preference Among Twitter Followers. In *Proceedings of the 29th International Conference on Human Factors in Computing Systems*, Vancouver, British Columbia, Canada, 2011. URL <http://portal.acm.org/citation.cfm?id=1979106>.

Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning Influence Probabilities In Social Networks. In *Proceedings of the 3th ACM International Conference on Web Search and Data Mining*, pages 241–250, New York, New York, USA, 2010. ACM Press. ISBN 9781605588896.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, page 27, New York, New York, USA, 2010. ACM Press. ISBN 9781450302456. URL <http://portal.acm.org/citation.cfm?doid=1866307.1866311>.

Pedro H Calais Guerra, Adriano Veloso, Wagner Meira Jr., and Virgílio Almeida. From bias to opinion : A transfer-learning approach to real-time sentiment analysis. *Machine Learning*, pages 150–158, 2011.

Aric Hagberg, Dan Schult, and Pieter Swart. Networkx. High productivity software for complex networks. <https://networkx.lanl.gov/>, 2010. URL <https://networkx.lanl.gov/>.

Thorsten Hennig-Thurau, Kevin P Gwinner, Gianfranco Walsh, and Dwayne D Gremmler. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1):38–52, 2004. URL <http://linkinghub.elsevier.com/retrieve/pii/S1094996804700961>.

Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, Toronto, Ontario, Canada, June 2010. ACM Request Permissions. URL <http://portal.acm.org/citation.cfm?id=1810617.1810647&coll=DL&dl=GUIDE&CFID=23680697&CFTOKEN=22562609>.

- Bernardo A Huberman, Daniel M Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. Technical report, 2008. URL [http://scholar.google.com/scholar?q=related:vXzQhrS8\\_G4J:scholar.google.com/&hl=en&num=30&as\\_sdt=0,5](http://scholar.google.com/scholar?q=related:vXzQhrS8_G4J:scholar.google.com/&hl=en&num=30&as_sdt=0,5).
- Gabriel Ishida. Mensurando Influência no Twitter: o índice Klout, May 2011. URL <http://www.dp6.com.br/mensurando-influencia-no-twitter-o-indice-klout>.
- Raj K. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley/Interscience, April 1991. ISBN 978-0-471-50336-1. URL <http://www.cse.wustl.edu/~jain/books/perfbook.htm>.
- Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, November 2009. URL <http://doi.wiley.com/10.1002/asi.21149>.
- Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June 2011. URL <http://portal.acm.org/citation.cfm?id=2002472.2002492&coll=DL&dl=GUIDE&CFID=61791103&CFTOKEN=31572846>.
- Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML'98*. Springer-Verlag, 1998. ISBN 3-540-64417-2.
- Elihu Katz, Paul Lazarsfeld, and Elmo Roper. *Personal influence: the part played by people in the flow of mass communications*. Transaction Publishers, 1955. URL <http://psycnet.apa.org/psycinfo/1956-05938-000>.
- Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the 1st Workshop on Online social Networks*, page 19, New York, New York, USA, 2008. ACM Press. ISBN 9781605581828.
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, page 591, Raleigh, North Carolina, USA, 2010. ACM

Press. ISBN 9781605587998. URL <http://portal.acm.org/citation.cfm?doid=1772690.1772751>.

Eun Sook Kwon and Yongjun Sung. Follow Me! Global Marketers' Twitter Use. *Journal of Interactive Advertising*, 12:4–16, 2011. URL <http://jiad.org/download?p=149>.

Paul Felix Lazarsfeld, Berelson Berelson, and Hazel Gaudet. *The people's choice: how the voter makes up his mind in a presidential campaign*. Columbia Univ. Press, 1948.

Alex Leavitt, Evan Burchard, David Fisher, and Sam Gilbert. The influentials: New approaches for analyzing influence on twitter. Technical report, 2009. URL [http://scholar.google.com/scholar?q=related:Pm29kDGLGs8J:scholar.google.com/&hl=en&num=30&as\\_sdt=0,5](http://scholar.google.com/scholar?q=related:Pm29kDGLGs8J:scholar.google.com/&hl=en&num=30&as_sdt=0,5).

Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina, USA, April 2010a. ACM. URL <http://portal.acm.org/citation.cfm?id=1772842>.

Jumin Lee, Do-Hyung Park, and Ingoo Han. The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications*, 7(3):341–352, 2008. URL <http://linkinghub.elsevier.com/retrieve/pii/S1567422307000415>.

Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Request Permissions, July 2010b. URL <http://portal.acm.org/citation.cfm?id=1835522>.

Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 199–208. IKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, October 2010. ISBN 978-1-4503-0099-5. URL <http://portal.acm.org/citation.cfm?id=1871437.1871467>.

Michael Mathioudakis, Nick Koudas, and Peter Marbach. Early online identification of attention gathering items in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 301–310, New York, NY,

- USA, 2010. ACM. ISBN 978-1-60558-889-6. URL <http://doi.acm.org/10.1145/1718487.1718525>.
- Robert K. Merton. Patterns of influence: Local and cosmopolitan influentials: Social theory and social structure. pages 441–74, 1968.
- T. Mitchell. *Machine Learning (Mcgraw-Hill International Edit)*. McGraw-Hill Education (ISE Editions), 1st edition, October 1997a. ISBN 0071154671. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0071154671>.
- Tom M Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1 edition, March 1997b. ISBN 0070428077. URL <http://www.amazon.com/Machine-Learning-Tom-M-Mitchell/dp/0070428077%3FSubscriptionId%3D1V7VTJ4HA4MFT9XBJ1R2%26tag%3Dmekentosjcom-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0070428077>.
- Richard W Mizerski. An attribution explanation of the disproportionate influence of unfavorable information. *Journal of Consumer Research*, 9(3):301–10, 1982.
- Brendan O’Connor and Ramnath Balasubramanyan. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, D.C., USA, 2010. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1536/1842>.
- Aditya Pal and Scott Counts. Identifying Topical Authorities in Microblogs. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 45–10, New York, New York, USA, 2011. ACM Press. ISBN 9781450304931.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, May 2002. URL <http://www.citebase.org/abstract?id=324042>.
- Carolyn Penner. #numbers. *Twitter Blog*. <http://blog.twitter.com/2011/03/numbers.html>, March 2011. URL <http://blog.twitter.com/2011/03/numbers.html>.
- Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, 2009. URL <http://portal.acm.org/citation.cfm?id=1639794>.

- Ronaldo C. Prati and Maria C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6, 2004.
- Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*. Association for Computational Linguistics, June 2005. URL <http://portal.acm.org/citation.cfm?id=1628969>.
- Sandy Ressler. *Perspectives on electronic publishing: standards, solutions, and more*. 1993. ISBN 0-13-287491-1.
- Thomas S Robertson and James H Myers. Personality correlates of opinion leadership and innovative buying behavior. *Journal of Marketing Research*, 6(2):164–168, 1969.
- Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and Passivity in Social Media. In *Proceedings of the 20th International Conference on World Wide Web*, page 113, New York, New York, USA, 2011. ACM Press. ISBN 9781450306379.
- Britta Ruhnau. Eigenvector-centrality – a node-centrality? *Social Networks*, 22(4): 357–365, October 2000.
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, April 2010. URL <http://portal.acm.org/citation.cfm?id=1772777>.
- Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM Request Permissions, November 2009. URL <http://portal.acm.org/citation.cfm?id=1653771.1653781&coll=DL&dl=GUIDE&CFID=23680697&CFTOKEN=22562609>.
- Neil Savage. Twitter as medium and message. *Communications of ACM*, 54(3): 18, March 2011. URL <http://portal.acm.org/citation.cfm?doid=1897852.1897860>.
- Leon Schiffman and Leslie Kanuk. *Consumer Behavior (7th Edition)*. Pearson Education, 7th edition, 1999. ISBN 0130841293. URL <http://www.worldcat.org/isbn/0130841293>.

Maggie Shiels. Twitter co-funder Jack Dorsey rejoins company. *BBC News*: <http://www.bbc.co.uk/news/business-12889048>, May 2011. URL <http://www.bbc.co.uk/news/business-12889048>.

Adrian J Slywotzky and Benson P Shapiro. Leveraging to beat the odds: The new marketing mind-set. *Harvard Business Review*, 71(5):97–107, 1993.

Michael Speriosu, Nikita Suda, Sid Upadhyay, and Jason Baldrige. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. In *Proceedings of EMNLP 2011. Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, 2011. URL <http://www.aclweb.org/anthology-new/W/W11/W11-22.pdf#page=63>.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short Text Classification in Twitter to Improve Information Filtering. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 841. ACM Press, 2010. ISBN 9781450301534.

The Nielsen Company. Global Advertising: Consumers Trust Real Friends and Virtual Strangers the Most, July 2009. URL <http://blog.nielsen.com/nielsenwire/consumer/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most/>.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2): 406–418, December 2010. URL <http://doi.wiley.com/10.1002/asi.21462>.

Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 565–574, New York, New York, USA, 2011. ACM Press. ISBN 9781450304931.

Christophe Van den Bulte and Yogesh V. Joshi. New Product Diffusion with Influentials and Imitators. *Marketing Science*, 26(3), May 2007. URL <http://portal.acm.org/citation.cfm?id=1528706.1528714&coll=DL&dl=GUIDE&CFID=51494273&CFTOKEN=62165757>.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.



- Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994. URL [http://books.google.com/books?hl=en&lr=&id=CAm2DpIqRUIC&oi=fnd&pg=PR21&dq=social+network+analysis+methods+and+applications&ots=HuHls90zLa&sig=NUbLmqfmUZDAS\\_DA86bLOJ4E2K8](http://books.google.com/books?hl=en&lr=&id=CAm2DpIqRUIC&oi=fnd&pg=PR21&dq=social+network+analysis+methods+and+applications&ots=HuHls90zLa&sig=NUbLmqfmUZDAS_DA86bLOJ4E2K8).
- Duncan J. Watts and Peter Sheridan Dodds. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*, 2007.
- Kevin Weil. Measuring Tweets. *Twitter Blog*. <http://blog.twitter.com/2010/02/measuring-tweets.html>, 2010. URL <http://blog.twitter.com/2010/02/measuring-tweets.html>.
- Weka Machine Learning Project. Weka. URL <http://www.cs.waikato.ac.nz/~ml/weka>, 2009.
- Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. TwitterRank: finding topic-sensitive influential twitterers. In *Proceedings of the 3th ACM International Conference on Web Search and Data Mining*, New York, New York, USA, February 2010. ACM Request Permissions. URL <http://portal.acm.org/citation.cfm?id=1718520>.
- F. Dianne Lux Wigand. Twitter takes wing in government: diffusion, roles, and management. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*. Digital Government Society of North America, May 2010. URL <http://portal.acm.org/citation.cfm?id=1809889>.
- Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181:1138–1152, March 2011.
- Qiang Yang and Xindong Wu. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*, 5(4):597–604, 2006.
- Zicong Zhou, Roja Bandari, Joseph Kong, Hai Qian, and Vwani Roychowdhury. Information resonance on Twitter: watching Iran. In *Proceedings of the 1st Workshop on Social Media Analytics*. ACM Request Permissions, July 2010. URL <http://portal.acm.org/citation.cfm?id=1964858.1964875&coll=DL&dl=ACM&CFID=23490382&CFTOKEN=79407738>.