# Media professionals' opinions about interactive visualizations of political polarization during Brazilian presidential campaigns on Twitter

Caroline Q. Santos[*], Holisson S. da Cunha[*], Carlos Roberto G. Teixeira[†], Daniele R. de Souza[*], Roberto Tietzmann[†],
Isabel H. Manssour[*], Milene S. Silveira[*], Duncan D. A. Ruiz[*], Marcelo Träsel[†], Rodrigo C. Barros[*]
[*] *PUCRS, Faculdade de Informática*
[†] *PUCRS, Faculdade de Comunicação Social*
*Porto Alegre, Brazil*
{*caroline.queiroz, holisson.cunha, carlos.teixeira, daniele.souza*}*@acad.pucrs.br,*
{*rtietz, isabel.manssour, milene.silveira, duncan.ruiz, marcelo.trasel, rodrigo.barros*}*@pucrs.br*

## Abstract

Interactive data visualization techniques are an important way to obtain information from large datasets. Data journalism is an emerging area that strongly makes use of such techniques. In this work we investigate the relationship between journalists (and media professionals) in their job routine and data visualization, with the main goal of understanding if these professionals know and use data visualization tools in their job context, as well as if they consider these resources to be important. For this, we present the results of a survey made with journalists and media professionals to analyze how interactive visualizations could help them to get insight or knowledge of such data, and if their use may improve and support these professionals' activities. The results indicate that visualization and data analysis tools are still not easily accessible by those professionals, and therefore still less influential than they could be. However, most participants considered data visualization a valuable resource in their news production routines. As a contribution, we also identified positive points and understanding gaps of visualizations, as well as the perception of journalists and media professionals about getting information from data visualization.

## I. Introduction

Data journalism (or computational journalism) is the application of computing and computational thinking to the usual activities of journalism [1], [2], [3]. This creates "new possibilities due to the combination of the common-sense 'nose for news' and the ability to tell a compelling story with the sheer scale and range of digital information now available" [4]. According to Diakopoulos [2], this field draws on some sub-fields of computer science, as information retrieval, artificial intelligence, language processing, visualization, personalization, among others, as well as aspects of social computing. It does not imply an obsolescence of traditional journalism routines, but it is an addition to them.

Data visualization has become a popular medium for journalistic storytelling and professionals have been increasingly using it in news stories to discuss complex issues including elections, economy and global health [5], [6], [7], [8]. In this sense, the visualization operates as a means to find relevant evidence that cannot be captured by a photograph or regular news reporting.

In this paper we are concerned with the journalists and media professionals' perceptions about data visualization. Through the design and evaluation of four interactive visualizations developed by our research group - specially focusing on the target audience - we are exploring prototypical news analysis in response to social media data. Our goal is to understand if journalists and media professionals know and use data visualization tools in their job context, as well as if they consider these resources to be important.

In a preliminary study about user behavior and sentiments in social networks [9], we used tweets collected during the World Cup in 2014, and we created some visualizations about the users' sentiment regarding the Brazil vs. Germany match, based on bar and line graphs. In order to analyze the visualizations' adequacy in improving understanding about that episode, they were presented to journalists and Journalism students in two focus groups, and their analysis and perceptions helped us to evolve our research in this area.

Taking advantage of the focus groups' contributions [9] and continuing our research, our attention remained on Twitter and how data visualization could support journalism professionals to obtain information. The timely context this time were the Brazilian presidential elections in 2014. We collected and processed tweets during the presidential campaign and we developed interactive visualizations to show the voters' sentiments about the candidates. They were presented to journalists and media professionals through an online survey and, in this paper, we discuss its results. Our main contributions include:

- identifying positive points in the developed interactive visualizations;
- identifying some understanding gaps in the developed interactive visualizations;
- discussing machine learning techniques used and their benefits to this study; and
- analyzing the participants' perception of getting infor-

URI: http://hdl.handle.net/10125/41383
ISBN: 978-0-9981331-0-2

mation from data visualizations.

The remainder of this paper is organized as follows. Some background concepts and related work are presented in sections II and III. Sections IV and V describe the used methodology and the obtained results, respectively. Section VI contains a discussion about our work. Finally, we present our conclusions and goals for future research in section VII.

## II. Background

This section presents basics of data-driven journalism, visualization and analysis of social media data to level the concepts with readers who are new to this area.

### A. Data-driven journalism

Data-Driven Journalism (DDJ) refers to the journalistic professional practices that involve the use of large datasets and computational resources to develop more complete and reasoned news stories [4]. Those practices can be categorized in three types [10]: Computer-Assisted Reporting (CAR), Data Journalism (DJ), and Computational Journalism (CJ). CAR, originally called Precision Journalism, is close to investigative reporting and has it roots in social science-based statistical methods, emphasizing the data gathering and statistical analysis and more general computer-based information retrieval skills, such as online and archival research, and email interviews. DJ may be considered the contemporary CAR and also the term of preference for journalism based on data analysis and the presentation of such examinations. Thus, it is the process of "obtaining, reporting on, curating and publishing data in the public interest" [11]. CJ, in its turn, includes at the same time CAR and DJ, as "the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism" [12]. In other words, it is the application of computing and computational thinking to the practices of information gathering, sense-making, and information presentation, rather than the journalistic use of data or social science methods more generally speaking [2]. CJ works as a flow from data to information to knowledge [3], suggesting relevant questions more than simple answers [1].

### B. Data visualization

Visualization can be defined as communication of information using graphical representations [13]. It transforms data, information, and knowledge into a form in which the human visual system perceives the embedded information on it [14]. According to Ward [13], visualization is important because *"we are visual beings who use sight as one of our key senses for information understanding"*. Therefore, the goal of visualization is to aid the understanding of data by leveraging the human visual system ability to recognize patterns, spot trends, and identify outliers [15]. If visualizations are well-designed, they can improve data comprehension and give viewers an immediate and profound impression, besides cutting through the clutter of a complex story to get right to the point [4].

The research in this area has advanced new forms of data collection, manipulation and interaction, mingling with other fields and processes. In this context visualization has been a pillar of the data journalism field. According to Gray et al. [4], in the reporting phase, visualizations can help journalists identify themes, questions, trends and unusual deviations, find typical examples and even suggest gaps and omissions in reporting. Furthermore, visualizations also play multiple roles in publishing as they illustrate a point made in a story in a more compelling way, remove unnecessarily technical information, and suggest transparency about journalists' reporting process to the readers.

### C. Social Media Data Analysis

Professionals who work with data (as journalists, researchers and analysts) have a multitude of computational tools available to assist the collection, cleaning, analysis and presentation of data. Examples of these tools such as Google Sheets, Web Scraper, OpenRefine, Infogram, Quadrigam, Google Analytics, Tableau, Gephi, etc. are abundant. However, according to Brooks [16], incompatibilities and design limitations may demand expert skills from these professionals, such as performing extensive adaptations, finding fragile workarounds, or context shifts that can hinder their progress. This may restrict individuals' participation in this emerging research area.

According to Heer and Shneiderman [17], analysis requires contextualized human judgment regarding the domain specific significance of clusters, trends, and outliers discovered in data. *"How do they organize their information for analysis? Which computational tools do they apply? How do they collaborate with others? What are their analysis products?"* Those are questions that Chin Jr. et al. [18] presented to instigate research, development, and deployment of information technologies to support intelligence analysis.

Some researchers have conducted studies to understand data analysis practices in several domains, as intelligence analysis [18], but, as well as other authors [16], [19], [20], we have focused on the data analysis practices of social scientists working with social media data. In this study, specifically, we have focused on the data analysis practices of journalists and media professionals working with social media data to enhance reporting and understanding.

Thus, researchers on social media data face some methodological and technical barriers and issues about how research with online social data should be done, ensuring, e.g., validity, ethics, and reproducibility. So, the design of data analysis tools is an interdisciplinary challenge that requires understanding the domain in which the data analyst works, and other technical fields [16], [21] such as media and journalism.

## III. Related Work

Due to popularity of social networks and their ability to generate large volumes of opinionated data, several studies have explored computational techniques involved in data analysis process [19], [20], [22], [23]. Related to data preprocessing, machine learning techniques have been used to extract the sentiment of the online public (similarly to what was done in this study). Sentiment analysis tasks usually collect and store data in natural language, requiring preprocessing efforts to improve data quality and achieve better results. For instance, Pak *et* al. [24] explored strategies trying to reduce the dimensionality and remove irrelevant terms considered in the classification. Wang *et* al. [25] developed a system to analyze the sentiment of voters regarding the candidates for the US presidency in 2012, in which the preprocessing used an approach of tokenization, identification of emoticons and exclamation points. The text was represented by unigrams and learning through the Naïve Bayes classifier. Similar to Pak's study, in this paper we explore such techniques to reduce the dimensionality and increase the representation of features.

Our work is inspired by Diakopoulos et al. [19], who share our focus on supporting the job of journalists and social media professionals through the development of interactive visualizations. In their work, the authors developed a visual analytic tool to make the social media response to events more amenable to journalistic investigation and sensemaking. The focus was on designing a tool that is able to direct attention to the pieces of information that may be most interesting journalistically speaking, as well as to schematize visual representations in ways that enable improved journalistic inquiry. Our work focused on a previous stage to understand journalists' perception about getting information from data visualizations and how they could use them in their jobs. As the authors [19], we also conduct a user study to collect the opinion of the target audience for future development of new visualizations and evaluation in the context of journalistic sensemaking.

Another related work [22] presents a tool for creating customized visualizations without the need of coding. It supports an expressive range of visualization designs and exports visualization components that can extend existing typologies. The authors evaluate tool's expressivity and accessibility through examples that are difficult to construct with existing interactive tools. They also carry out a study with journalists and a first-use study with visualization designers and journalists. Our work builds on the ideas presented by these authors [22] in several important ways, as in merging a range of visualization techniques (see IV-C) and in investigating journalists' preferences for getting information from data visualizations.

Regattieri et al. [23] proposed a visualization technique to reveal behaviours and discourses in a large dataset from Twitter. They created an interactive model that allows to identify part and whole pattern relationships, constant with the three principles of information visualization: overview first, zoom and filter, and then details on demand. The authors analyze relationships in a social network in order to create a visualization ready to support users when telling a story with data. The tool was considered appropriate to journalists to visualize news and events, since it allows them to use data to communicate something everyone can understand and relate to real events. Our focus was not on creating visualizations about relationships in social network, but we share the same goal: to ensure that visualizations of massive amounts of information are increasingly useful and accessible to journalists and social media professionals.

## IV. Research Description

This section presents the developed research, including a methodology overview, an explanation about data gathering and the preprocessing stage, and a description of the proposed interactive visualizations.

### A. Methodology Overview

In a preliminary study [9], we started the process of learning how visualization techniques could help media students and journalists to enhance their understanding about user behavior and sentiments in social networks. We used tweets collected during the World Cup in 2014, and we generated bar and line graph visualizations related to the users' sentiment in the Brazil vs. Germany match. In order to analyze visualization adequacy in improving understanding about the episode, these visualizations were presented to journalists and Journalism students in two focus groups. Taking advantage of their feedback, we continued investigating how data visualization can support journalism professionals to get more information about an event. We followed the same methodology of our preliminary study, a descriptive research [26], where we started generating a description about the opinions of journalists and media professionals concerning data visualization and how much it could assist in their jobs.

Thus, we conducted a qualitative study through the application of an online survey to journalists and media professionals working with social media data. The main question that guided the survey was: *how information obtained from the presented visualizations may support journalists' activities?* To explore data visualization possibilities, we collected and processed tweets during the presidential campaign to Brazilian election in 2014, and these data were presented to the participants through four interactive visualizations.

The survey was written in Portuguese, comprising 29 questions divided into three sections: the first section with the first twelve questions about the participants' profile and their experiences with social networks. The second section presented questions (from question 13 until 24) to collect information about the four developed interactive data visualizations and the participants' perceptions about them. There was one closed question and three open-ended questions to each visualization. The visualizations were presented by means of screen recordings with voiceovers. Finally, the third section (from question 25 until 29) asked about the participants' opinions regarding data visualization, in general. All questions from this last section were open-ended. The survey intended outcome was to synthesize the opinions about the elements presented in an interactive visualization. The survey was distributed by email to people who had been previously researched as having a profile related to our target audience, groups (on social networks), and lists of professionals who work with social media data.

In the following sections, the data gathering and processing steps, and the visualizations created for this study are described. The analysis of answers from the open questions is the main component of our Results section.

### *B. Data Input and Processing*

In recent years, Twitter has become a valuable tool for communication and sharing opinionated content, addressing various topics of interest to society. This paper analyzes a database of tweets from the second round of Brazilian presidential election in 2014, collected between 6 and 26 of October 2014, using as query terms the names of the candidates Dilma Rousseff (the standing president, attempting reelection) and Aécio Neves (opposition).

Data were collected through the free version of the Twitter Streaming API, achieving a total of 150,687 tweets about Aécio and 177,407 regarding Dilma. In order to identify the sentiment of the online public and analyze changes in user polarization, it was used a classification algorithm, so that data was categorized according to their polarity, whether positive or negative. For this, Machine Learning (ML) techniques were applied, generating a decision model using pre-labeled data (training data) to classify data without labels. The goal through ML techniques was to automatically find patterns in large volumes of data, extracting knowledge hither to represented in an implicit form.

In Sentiment Analysis (SA), the classification problem was divided into two steps: (1) learning a decision model using previously labeled training; and (2) predicting data polarity from the decision model. The acquisition of the training set was performed using tweets containing emoticons and hashtags with negative or positive polarity. During the elections, several hashtags were created by Twitter users to describe a specific topic. Thus, these 30 hashtags were manually labeled, considering the sentiment expressed toward each candidate. The decision on the sentiment of a tweet was carried out in a simplified manner: if a positive hashtag is present in a tweet, this tweet is then labeled as positive, whereas if the hashtag is negative, then the tweet is labeled as negative. In addition, negative hashtags regarding candidate X and mentions to the name of candidate Y were considered positive for Y. Through this strategy, 55,000 tweets labeled for candidate Aécio and 73,000 for candidate Dilma were obtained. The remainder of the data was labeled using the generated decision model. Traditional ML algorithms are usually not able to directly process textual data written in natural language. Therefore, during the preprocessing task, it was required to convert messages in one structured representation model capable of being manipulated and processed by the classification algorithm.

In SA, the most common way to generate a structured model is turning it into a sparse number vector, suitable for processing through algebraic operations. This representation, called bag-of-words (BOW), refers to a n-dimensional data space, where n is the number of attributes, and each dimension is represented by an attribute and their respective weight. BOW represents each message as a set of independent attributes, disregarding the order in which the terms originally occur in the text [27]. Moreover, short and informal texts represent an important challenge in Sentiment Analysis. On Twitter, messages are limited to 140 characters and are constantly used to share feelings and subjective expressions. Because of this limitation, it is common to find abbreviations and slang, as well as grammatical errors and other textual structures that are characteristic of social networks [28]. Due to the importance of text treatment in the classification process, in this paper we explore a set of techniques aiming at identifying the subjectivity of informal texts to achieve better results in the process of identifying sentiments. The techniques applied are described below.

*Filtering:* in order to reduce the dimensionality and eliminate noise in the data, the following text elements have been removed: 1) web links, such as sequences of characters that start with "http" or "www". 2) Query terms used to collect tweets to avoid their influence over the tweets' classification. 3) Usernames and mentions to other users in the following format "@username", and 4) special characters, such as "RT" and non alpha-numeric elements.

*Tokenization:* it uses rules and regular expressions to split the text into tokens (basic units of a language). These tokens are usually words and other elements used in language;

*Stopword removal:* stopwords are common words that usually do not contribute to the analysis, such as: there, are, you, and we. In this work, a dictionary of common words was used for the Portuguese language.

*Spell check:* tweets frequently contain orthographical mistakes that contribute to a dimensionality inflation and affect data processing. To reduce this problem we used a natural language library called textBlob that automatically performs text spell-checking.

*Lower casing:* it is common to find different structure variations in texts shared throughout Twitter (e.g., "DAta

Table I
EXPERIMENTS RESULTS

| Candidate | Precision | Recall | F-measure |
|---|---|---|---|
| Dilma | 74.00 | 66.45 | 70.02 |
| Aécio | 87.11 | 86.81 | 86.96 |

MininG" and "DaTa MiNiNg"). In order to establish a standard in the preprocessing and ensure consistency in the text treatment, all characters were converted to lowercase.

*Negation identification:* bigrams were built to identify negative pieces of polarity on tweets. To build the bigrams, we used the presence of denial such as: "not, no, never, don't", and, for each denial found, bigrams were formed joining the denial with the consequent token as in "not good". As per [24], [29], this process improves accuracy, since the denial presence presents an important text structure in Sentiment Analysis problems.

The classifier used was Multinomial Naïve Bayes, a traditional statistical classifier based on Bayes theory which is often used in text mining problems. This approach considers a document as a bag of words. For each class p(w—c), the probability of observing word w given the occurrence of c classes is estimated by training data, simply by calculating the relative frequency of each word in the training set. The classifier also requires the probability a *priori* P(c) [30].

In order to reduce the impact of the variability of the data, it was applied cross-validation using 10 folds on the training data for each candidate considered. To evaluate the predictive quality of the ML algorithm, 3 standard measures in sentiment analysis were used : precision, recall, F-measure. Table I shows the results.

Considering the preprocessing techniques used, the features set generated contains words, hashtags and bigrams with negated words. In addition to reducing the dimensionality of the data, the preprocessing techniques improve the predictive quality on sentiment analysis in 2%.

### C. Interactive Visualizations

We created four interactive visualizations to use with the tweets collected as explained in section IV-B. Visualizations were created in Portuguese, but, for the sake of English readers' comprehension, the most relevant texts and labels were translated. They have been made available on Youtube[1] with narration in Portuguese.

We used D3.js[2] and, basically, all of them have a similar minimal structure, as shown in figure 1: 1 and 2 - a suite of auxiliary controls, 3 and 4 - timeline, and D3's brush component (to implement focus + context zooming), respectively.

To facilitate discussions, visualizations were called as:

- Graphic 1: interactive line chart (figure 2).
- Graphic 2: interactive multiline chart (figure 3).
- Graphic 3: interactive circle chart (figure 4).
- Graphic 4: interactive line chart integrated with news (figure 7).

In all visualizations, x-axis represents time over the days between the first and second round of the election. In Graphic 1 (figures 1 and 2), y-axis represents positive (top of the y-axis) and negative (bottom of the y-axis) sentiments. The orange line represents sentiments about candidate Dilma, and the blue line represents sentiments about candidate Aécio. By the graphic, we can notice that there are more positive tweets about Dilma than positive tweets about Aécio. An important feature is the brush component, that allows focusing on a specific time interval and expanding the graphic area (context zooming). Figure 2 shows the interactive line chart (Graphic 1) after focus + context zooming. This visualization allows sentiments' rhythm analysis over time.

Graphic 2 is an interactive multiline chart in which y-axis represents the amount of collected positive and negative tweets. The suite of auxiliary controls allows choosing among five frequency times, varying from 5, 10, 15 and 30 minutes, to 1 hour. The standard representation in this visualization is 30 minutes, but users can choose any of the others. There is not any candidate identification in this graphic, and the amount refers to the sum of positive and negative tweets from both candidates, Dilma and Aécio. It is possible to notice that lower peaks happened during the dawn, and there are spikes when debates between the candidates were held on television.

Graphic 3 is an interactive circle chart (figure 4), which shows the amount of tweets from both candidates, Dilma and Aécio. The sentiment is indicated by color: green for positive and red for negative. The circle size indicates the amount of tweets, and the larger circles at that moment were on the top of the graphic. In this visualization, the suite of auxiliary controls allows choosing a visual representation by subject matter, thus tweets were separated by candidate (see figure 5). Most tweets from candidate Dilma were positive, and from candidate Aécio were negative, as shown in this visualization. Circles about Dilma appear on top because there were more tweets from this candidate in the collected database, as mentioned in section IV-B. By positioning the mouse over a circle, the equivalent amount of tweets is shown, and, by clicking on any circle, a new visual representation is opened with tweets from that moment, separated in two groups (positive and negative), represented by new circles. Each new circle represents a user and the circle size refers to the amount of followers of that user. After this is done, by positioning the mouse over a circle, the text message from that tweet is shown (see figure 6).

The last visualization (Graphic 4) is an integrated interactive line chart with news (figure 7), an idea raised in our preliminary study [9]. The interactive line chart is the

[1] https://goo.gl/31HqEC
[2] https://d3js.org/

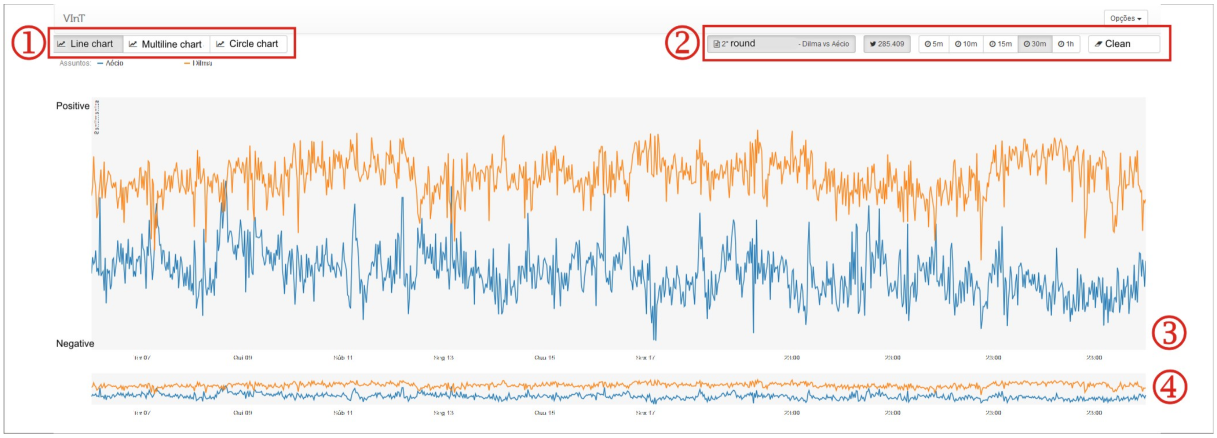


Figure 1. Visualizations' general structure with selection buttons on top.

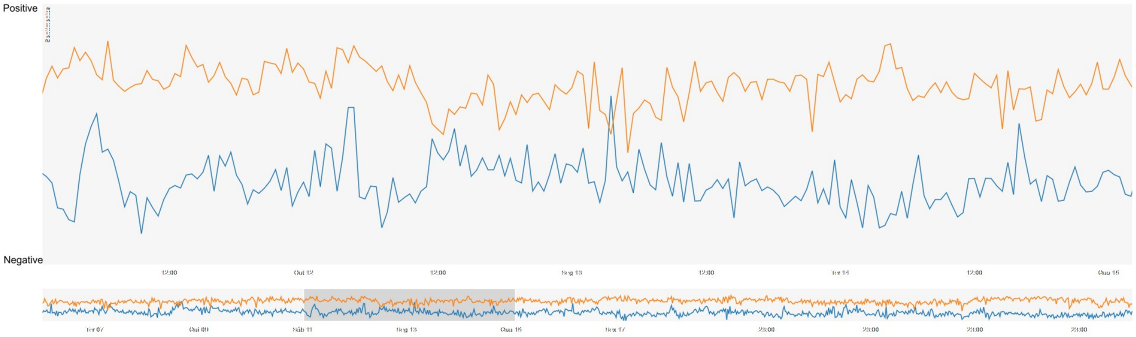


Figure 2. Interactive line chart after brushing - focus and context zooming.

same as Graphic 1, but with aggregated news. The news timeline was built using a tool called Timeline.js[3] with news collected during the same period of the tweets used in the visualizations. Thus, in this visualization we have the same line chart of figure 2, but on the top we have news (also collected from Twitter) about this subject. Regarding the news timeline specifically, by positioning the mouse over a piece of news, a screen area appears showing the original tweet. If we click over the tweet, that post on Twitter is opened in a new tab on the browser and users can see more details about it. This kind of visualization allows, for example, to speculate or create hypotheses about moments of high peaks in the line chart and the news posted in that moment.

## V. Results

After the visualizations' development, an online survey was undertaken with journalists and media professionals close to the field of social media data analysis. A total of 50 people answered our survey, in which 29 were male and 21 were female. 45 participants have accounts on Twitter and 37 participants use this medium as source of information. 74% of the participants (37) are graduated in journalism and 30 of them have already used infographics or data visualizations in their job activities. The participants are from different states of Brazil and were contacted as described in IV-A.

From the participants' opinion about the visualizations presented, we created four categories: criticism, compliment, contribution and impartiality. The interactive line chart (figure 2) was considered easy to understand. When answering about other possible visualizations, the bar and column chart were the most requested. When the question focused on sentiment analysis, one of the participants said: *"I think the question on the sentiments is very difficult. Positive and negative feelings, for me, are extremely complex, because of the lack of references to issues such as irony and sarcasm (positive or negative?). The references' question is clear in the line graphic, but the sentiments question isn't (besides being somewhat questionable)"*.

The research execution was also addressed. One of the participants considered it could be relevant to include an explanation about the process of collecting data and creating visualizations before the graphics were presented. Another factor to be considered was the possibility of splitting the graph into smaller parts and subdivisions, organizing them

[3] https://timeline.knightlab.com/

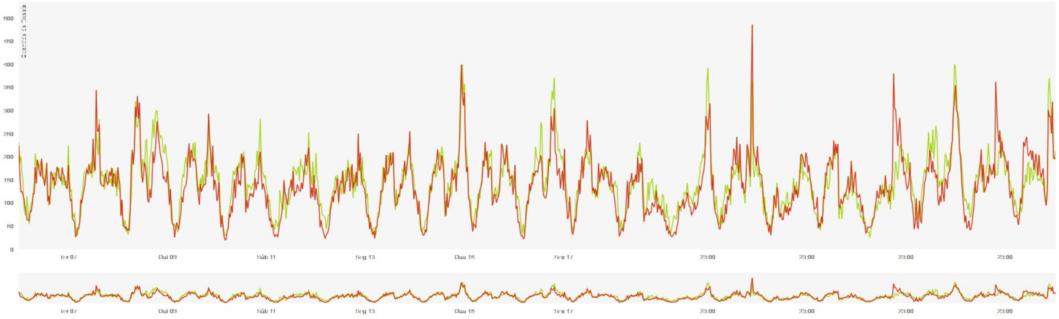


Figure 3. Interactive multiline chart representing the amount of election tweets organized by time.

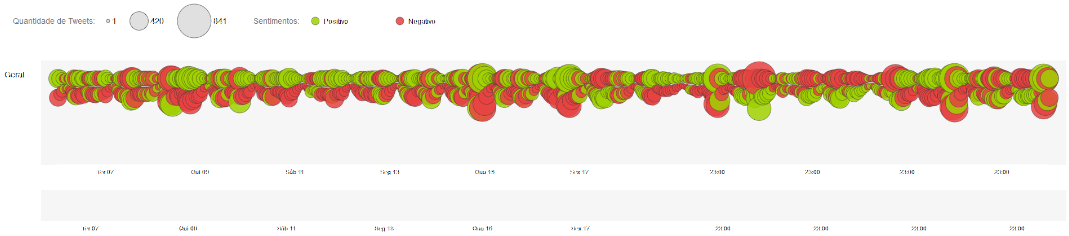


Figure 4. Interactive circle chart.

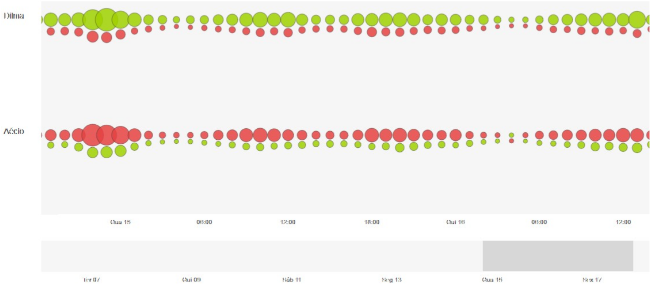
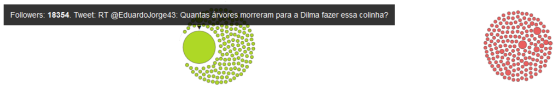


Figure 5. Interactive circle chat separated by subject matter.



Followers: **18354**. Tweet: RT @EduardoJorge43: Quantas árvores morreram para a Dilma fazer essa colinha?

Figure 6. Circles of unified groups of tweets by sentiment.

in time intervals, for example.

Beyond the sentiments' analysis, other possible indicators of positive, negative or neutral interpretations were raised: references, region and location. Another frequently addressed topic was the expansion of the data sources, suggesting the production of graphics from datasets collected on other social networks such as Facebook and Instagram. Alongside with that suggestion, the collection of markers and highlights typical of these networks (such as the likes and the number of comments, for example) was often proposed as a relevant step forward. Another possible future implementation was the construction of a visual relation of the collected Twitter data with external vote intention surveys, since a possible relation between both was suspected.

The interactive multiline chart (figure 3) was also considered easy to understand. In spite of that, the difficulty to discern the two lines shown together was mentioned. As a response to this, the use of visualizations with bars and columns were suggested again. The subdivision into smaller graphics also raised once more. One of the participants considered that the graphics *"journalistic terms, have little informative value."*. The possibility to sort the visualization by regions and amount of mentions were requested. Once again, the collection of data from other social networks (such as, again, Facebook and Instagram) was raised. It has also been proposed the construction of a *"visualization with journalistically relevant data, according to the content (themes) with feelings crossing (negative/ positive)"*, a goal considered to have been partially achieved in this figure.

In contrast, the interactive circle chart (figure 6) was considered hard to understand. When asked about visualization models, this chart was the one that received more criticism and viewing difficulties: *"I found it to be a mess, I can't identify anything"*; *"This graphic seems more confusing than the previous"*; *"It's more difficult to view and it's so confuse"*; *"...as it is, it would not help me to come to any conclusion"*. The only positive answers identified the possibility to expand each circle and view all its forming tweets grouped and expansible for individual reading.

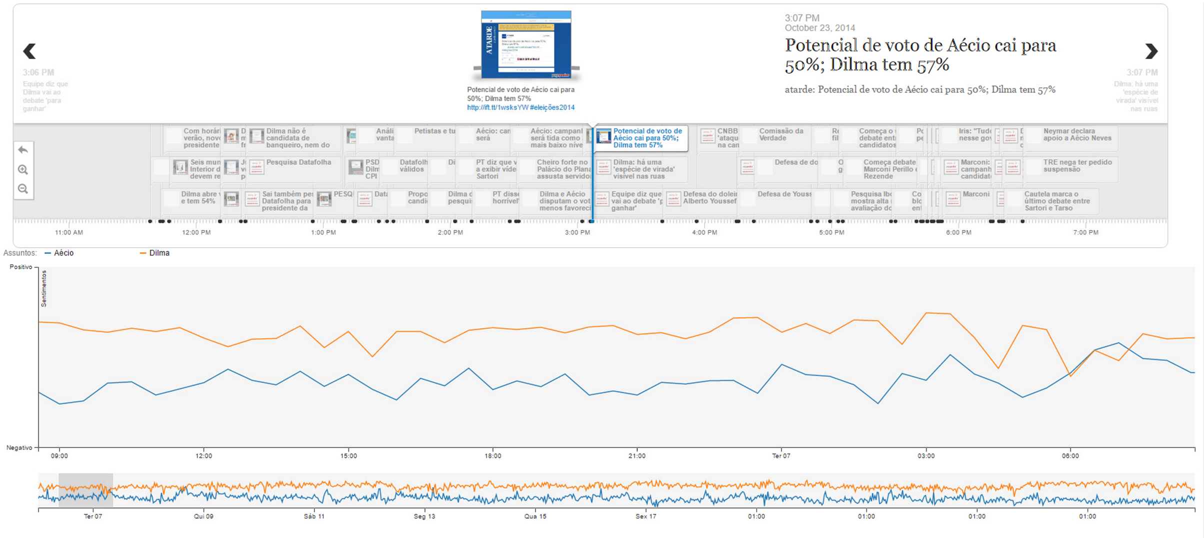

3:06 PM
Equipe diz que Dilma vai ao debate 'para ganhar'
Potencial de voto de Aécio cai para 50%; Dilma tem 57%
http://ift.tt/1wsksYW #eleições2014
3:07 PM
October 23, 2014
Potencial de voto de Aécio cai para 50%; Dilma tem 57%
atarde: Potencial de voto de Aécio cai para 50%; Dilma tem 57%
3:07 PM
Dilma: há uma 'espécie de virada' visível nas ruas
11:00 AM
12:00 PM
1:00 PM
2:00 PM
3:00 PM
4:00 PM
5:00 PM
6:00 PM
7:00 PM
Assuntos:
Aécio
Dilma
Positivo
Sentimentos
Negativo
09:00
12:00
15:00
18:00
21:00
Ter 07
03:00
06:00
Ter 07
Qui 09
Sáb 11
Seg 13
Qua 15
Sex 17
01:00
01:00
01:00
01:00

Figure 7. Integrated interactive line chart synchronized with election-themed news stories.

Finally, the interactive line chart integrated with news (figure 7) has been widely praised: *"It's a very good model, I can't think of a superior format"*; *"I loved the connection with the news"*. Subdivisions in the news area were suggested according to the publisher type (is it a regular newspaper online, a blog, etc.). Once again, the aggregation of results from other social networks and the segmentation by regions appeared as possible desired implementations.

About the data visualization influence in the journalistic practice, participants said that they think the use of data visualization techniques is very important, especially with regard to the understanding of a subject, making complex information more accessible to the public and accelerating the interpretation of them. Data visualizations were also highlighted as a relevant support to journalistic duties, whether in the establishment of facts and trends, in the preparation of a wider context to news stories or in the possible generation of new contents and narratives. One person spoke about the press office context, altogether different from the newsroom, in which data visualizations can serve as feedback to stakeholders' actions. Three replies addressed the question of feedback in: audience analysis, content qualification and public response to the content produced by a media. Participants also have spoken on the issue of the reader context and use of data visualizations to better understand information.

With regard to the subject types that could benefit from data views, eight people said that data views can benefit all or any subjects and sections. Among the people who answered the question, some specified subjects that could benefit more from data visualizations while others indicated news sections. Among the specified themes were elections, legislative measures, soccer and sports (hired and fired players, leagues and statistics, etc.), stock exchange index values, urban crime, among others. Categorizing these subjects in the news sections identified by the participants and creating some new ones, we have found that four stand out: politics, sports, economy, and security. Some participants also cited health, education, science and gastronomy sections and specific topics.

When asked about being able to generate agendas and future news stories from data visualization, participants agreed with the idea of agendas being generated from data views. One participant reported that, for this to be possible, cross-checking data tools need to be available in a simple format to reporters. Another person commented that what generates an agenda are the news criteria and the visualizations enrich the news, but do not replace it. There was also a response about data visualizations being the focal point of polemics and debates from predetermined polarized thought agendas before choosing which data would be used or not, possibly biasing its interpretation.

## VI. Discussion

In general, the visualizations were considered satisfactory. Participants claimed to be able to interpret a large volume of information in a more dynamic way through them. The main questions and criticisms that have arisen refer to the design graphics and the algorithmic classification of the feelings. This may have occurred because the survey has not addressed the methodology implemented to achieve such results. More than curating and visualizing the data itself, an explanation of how it was collected and treated is an important factor for those who analyze the graphics.

The visualization in form of lines (graphics 1 and 2) was the most praised. However, it was also widely suggested that the use of column and bar graphics could be able to present a better separation of layers of information, a relevant consideration according to Tufte [31]. Another important issue is the possibility of subdividing the graphic into smaller units. The idea of a single graph encompassing all information – even with interactive features such as the ability to focus on periods – did not appear to be enough to warrant adequate understanding. Still addressing subdivisions, besides the possibility of selecting a time interval and each one of the candidates, the selection by region and mentions were also suggested. The inclusion of other social networks, seeking a broader analysis, was considered relevant.

Graphic 3 was the most heavily criticized item of our visualizations. It is possible to conclude that users express the need to visualize a clearer organization of the information on separated layers and smaller chunks, an idea aligned to the concept of smallest effective difference as defined by Tufte [31] as the measure of data concentration and clarity on a graphic visualization. As the graphic elements did overlap, the more confusing the visualization became in the eye of our audience. However, a positive point recognized in this chart was the possibility of a bi-modal visualization: at first one sees all the grouped tweets and, when one hovers with the mouse, the tweets are presented individually. This is a feature that could be further explored on future implementations.

The participants showed a lot of interest in the possibility of a visual relationship representing the intersection of information: the inclusion of different social networks or the visualization of the interactions between the posts on social networks and news stories. Graphic 4 was highly praised, in spite of the suggestion to include more specific segmentation criteria having been also raised.

About the influence of data visualization in the journalistic practice, we perceived that data analysis tools still are not easily accessible (or known), and therefore still are not so much influential. One of answers can resume well this: *“when data tools are available in a simple and practical way for journalists, they will be interesting resources to generate agendas. The “Brazilian access to information” law has been valuable, but lack public investment to keep data more accessible”*.

Maybe we could make a connection with the introduction of CSCW (Computer-Supported Cooperative Work) in work processes in companies. As the CSCW community, visualization community should conduct more studies aimed at understanding how works takes place in practice. In line with Brooks [16], we believe that greater attention should be given to understanding data visualization users and the context in which data analysis work takes place.

Based on the responses obtained, three insights emerged for possible future works: the construction of tools aimed at crossing data and finding remarkable relations that are more accessible to journalists and content producers with a minimal familiarity with computer or information science; the choice of sports as an object of data collection to create visualizations; and develop visualizations about audience and content qualification.

## VII. Conclusions and Future Works

Data journalism is an area that benefits from interactive data visualization as a way to get insight or knowledge from large datasets, such as from social networks. Since social media tend to be increasingly integrated into journalistic practices, we presented the results of a research that analyzes how interactive visualizations from Twitter data can benefit journalists and media professionals. We worked with tweets gathered during the 2014 presidential elections in Brazil, generating different visualizations about the users' sentiment regarding the candidates. In relation to data preprocess, we used ML algorithms to identify the sentiment of Twitter users, and these techniques had high capacity to work with small error rates, achieving good results in this study.

We reported on a qualitative empirical study based on an online survey with 50 participants to understand if information obtained from the presented interactive visualizations may support job activities of journalists and media professionals. Through the design and evaluation of four interactive visualizations, we explored the domain of journalistic analysis in response to social media data, including implications to better design of interactive visualization tools. Our data analysis reveals that participants considered data visualizations a valuable resource in their job activities. However, it seems that data visualization (and data analysis tools) need to be better introduced in the working practices of these professionals, and their use need to be better studied.

Briefly, in this study we obtained feedback about the developed interactive visualizations; we identified understanding gaps in the developed interactive visualizations; we presented ML techniques used and their benefits to our study; and we obtained an overview of participants' perception about getting information from data visualizations.

We intend to continue to investigate and develop new visualization mechanisms for journalists and media professionals, to obtain information and knowledge from social media sources. A visualization that is being developed, e.g, intends to show the origin and use of different hashtags. Now we are gathering and processing tweets about the Brazilian President impeachment process and about the 2016 Olympic Games. In addition to the development and use of news interactive visualizations, based on guidelines or taxonomies as from Heer and Shneiderman [17], we started a cycle of semi-structured interviews. These interviews are being conducted only with people who are researching or working with data analysis from networks and online social media. The goal is to investigate data analysts' ability to identify their own needs about data visualization. The results will be used to deepen our understanding about their needs and to design new strategies to support their analysis. As future work, we also intend to make a qualitative study with users in which they can interact with the visualizations, so we can observe the way they interact with the tool to improve the user experience with it.

## Acknowledgement

This work was partially supported by PUCRS (Edital 07/2016 - Programa de Apoio a Integração entre Áreas/ PRAIAS). The authors acknowledge the collaboration of the students Jocines D. da Silveira, Lorenzo P. Leuck, Juvenal N. S. dos Santos Júnior and Maximilian N. S. Zorzetti.

## References

[1] S. Cohen, J. T. Hamilton, and F. Turner, "Computational journalism," *Commun. ACM*, vol. 54, no. 10, pp. 66–71, Oct. 2011.

[2] N. Diakopoulos, "A functional roadmap for innovation in computational journalism," *Nick Diakopoulos*, 2011.

[3] ——, "Cultivating the landscape of innovation in computational journalism," *Tow-Knight Center for Entrepreneurial Journalism*, 2012.

[4] J. Gray, L. Chambers, and L. Bounegru, *The data journalism handbook*. O'Reilly Media, Inc, 2012.

[5] J. Hullman and N. Diakopoulos, "Visualization rhetoric: Framing effects in narrative visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2231–2240, Dec 2011.

[6] A. Figueiras, "How to tell stories using visualization," in *Information Visualisation (IV), 2014 18th International Conference on*, July 2014, pp. 18–18.

[7] A. Satyanarayan and J. Heer, "Authoring narrative visualizations with ellipsis," in *Computer Graphics Forum*, vol. 33, no. 3, Wiley Online Library. EuroVis, 2014, pp. 361–370.

[8] B. Lee, N. Riche, P. Isenberg, and S. Carpendale, "More than telling a story: Transforming data into visually shared stories," *Computer Graphics and Applications, IEEE*, vol. 35, no. 5, pp. 84–90, Sept 2015.

[9] C. Q. Santos, R. Tietzmann, M. Träsel, S. M. W. Moraes, I. H. Manssour, and M. S. Silveira, "Can visualization techniques help journalists to deepen analysis of twitter data? exploring the "germany 7 x 1 brazil" case," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Jan 2016, pp. 1939–1948.

[10] M. Coddington, "Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting," *Digital Journalism*, pp. 331–348, 2015.

[11] J. Stray, "How the guardian is pioneering data journalism with free tools," *Nieman Journalism Lab*, 2010.

[12] J. T. Hamilton and F. Turner, "Accountability through algorithm: developing the field of computational journalism," in *A Center for Advanced Study in the Behavioral Sciences Summer Workshop. Duke University in association with Stanford University*, 2009, pp. 27–31.

[13] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA, USA: A. K. Peters, Ltd., 2010.

[14] N. Gershon and W. Page, "What storytelling can do for information visualization," *Commun. ACM*, vol. 44, no. 8, pp. 31–37, Aug. 2001.

[15] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo," *ACM Queue*, vol. 8, no. 5, pp. 20:20–20:30, May 2010.

[16] M. Brooks, "Human centered tools for analyzing online social data," Ph.D. dissertation, University of Washington, 2015.

[17] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," *Commun. ACM*, vol. 55, no. 4, pp. 45–54, Apr. 2012.

[18] G. Chin, Jr., O. A. Kuchar, and K. E. Wolf, "Exploring the analytical processes of intelligence analysts," in *Proceedings of the Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 11–20.

[19] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, Oct 2010, pp. 115–122.

[20] N. Diakopoulos, M. De Choudhury, and M. Naaman, "Finding and assessing social media information sources in the context of journalism," in *Proceedings of the Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 2451–2460.

[21] K. Crowston and K. Nahon, "Introduction to the digital and social media track," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, Jan 2015, pp. 1541–1541.

[22] A. Satyanarayan and J. Heer, "Lyra: An interactive visualization design environment," *Computer Graphics Forum*, vol. 33, no. 3, pp. 351–360, 2014.

[23] L. L. Regattieri, R. Chartier, J. Windsor, and G. Rockwell, "Tweetvis: Following twitter hashtags to support storytelling," in *CEUR Workshop Proceedings*, vol. 1210. ACM, 2014.

[24] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *Proceedings of the Seventh Conference on International Language Resources and Evaluation - LREC*, vol. 10, 2010, pp. 1320–1326.

[25] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 115–120.

[26] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. John Wiley & Sons, 2010.

[27] C. Aggarwal and C. Zhai, *Mining text data*. Springer, 2012.

[28] P. Priyanthan, B. Gokulakrishnan, T. Ragavan, N. Prasath, and A. S. Perera, "Opinion mining and sentiment analysis on a twitter data stream," *ICTer 2012*, 2012.

[29] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[30] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[31] E. R. Tufte and E. Weise Moeller, *Visual explanations: images and quantities, evidence and narrative*. Graphics Press Cheshire, CT, 1997, vol. 36.