2016 49th Hawaii International Conference on System Sciences

Can visualization techniques help journalists to deepen analysis of Twitter data? Exploring the "Germany 7 x 1 Brazil" case

Caroline Q. Santos*, Roberto Tietzmann†, Marcelo Träsel†, Sílvia M.W. Moraes*, Isabel H. Manssour*, Milene S. Silveira*
* *PUCRS, Faculdade de Informática*
† *PUCRS, Faculdade de Comunicação Social*
Porto Alegre, Brazil
caroline.queiroz@acad.pucrs.br,
{*rtietz, marcelo.trasel, silvia.moraes, isabel.manssour, milene.silveira*}*@pucrs.br*

Abstract

The use of social networks has increased steadily from the mid-2000s on, generating a large volume of data reflecting current affairs relevant to be analyzed using data visualization techniques that can emphasize important information otherwise concealed. Our main goal is to understand how visualization techniques can help media studies scholars and journalists create better understanding about user behavior and sentiments in social networks. This motivated an interdisciplinary research project where tweets sent during the World Cup in 2014 were collected and processed, generating visualizations about the users' sentiment in the Brazil vs. Germany game and hashtags most used by online Brazilians. In order to analyze the visualizations' adequacy in improving understanding about the episode, they were presented to journalists and students of Journalism in two focus groups, and their results analysis is presented in this paper.

I. Introduction

The use of social networks by growing masses of individuals has increased steadily from the mid-2000s on, generating large volumes of data reflecting current affairs and shifts in mood and opinion, and its understanding has become a strategic issue with applications in many scenarios. According to Russel [1], social networks are changing the way of life on and off the web, which allows the display technology and information about everything, or, as Levy [2] affirms, material and digital aspects of contemporary life have become intertwined more than only superimposed onto another. From this panorama arises an opportunity to interlace social web elements with computational techniques of data collection, preprocess and data visualization to support experts in analysis and knowledge extractions. Examples abound, such as the understanding of electoral campaigns and elections [3] [4] or inferring the political orientation of Twitter users [5] [6] [7], sports events [8], product and service satisfaction [9], among other studies related to massive events that could benefit from this. To better understand the data, techniques for enhancing data visualization and interaction are developed and improved.

According to Heer et al [10], 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. If visual representations are well-designed, they can replace intuitive cognitive calculations with simple perceptual inferences and improve comprehension, memory, and decision making. The challenge in visualization involves issues such as for any given data set, the number of visual encodings is extremely large [10].

Techniques for visualizing social media datasets have been used by several media outlets to illustrate people's behavior in one specific subject or event. In the 2014 World Cup in Brazil, data-driven journalistic visualizations were highlighted and widely shared in social media[1], and international news agencies[2]. The work involved in this usually requires several computational techniques (such as data and text mining, natural language processing, etc.) and complex analytical processes required to manipulate varied data sources. Besides that, reach a point of balance between the computational side of the process and the aesthetic side using tables, charts, colors and other visual features, could favor a good analysis and quicker understanding of such data. Many researchers [11] [12] [13] [10] have been involved with the study of the evolution of these techniques. In several situations, e.g., a simple line or bar chart were not good enough to translate the complexity of the data to a general audience and, therefore, our primary goal is to understand how visualization techniques can help media studies and journalism professionals and students to better understand user behavior (and also user sentiment) in social networks.

Our focus is on Twitter, investigating *how can data visualization support obtaining information that would be difficult to achieve and perceive otherwise?* For this, considering the

[1]http://t.co/jOoei6z69i and https://interactive.twitter.com/wcmentions/
[2]http://t.co/lvIF36lqzz and http://goo.gl/rPzfme

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.245

great repercussion in the national media over 2014 World Cup in Brazil, we collected and processed tweets during the World Cup in 2014 and we developed some visualization techniques related to Brazil vs. Germany semifinal[3]. The visualizations were presented to journalists and students of Journalism in two focus groups and, in this paper, we discuss its results.

The remaining of this paper is organized as follows. The next section presents the context relating journalism and data visualization in respect to this work. After this, some related works are presented. Then we describe the methodology used in our research by detailing the case study performed and its analysis. Finally, the last section presents closing comments and future works.

II. Journalism and Data Visualization

Data-driven journalism (DDJ) comprises several professional practices whose common thread is the use of databases as the main source of information for news production. DDJ practices involve Computer-Assisted Reporting, data visualization, infographics, creation and collection of databases, as well as policies of open-access to information and transparency of governments. Contemporaneously, other formulations used to refer to this professional specialty are "data journalism" [14] [15] or "computational journalism" [16] [17] [18] [19]. DDJ focuses on the production, processing and analysis of large amounts of data, to allow more efficient retrieval of information, investigative reporting from data sets, and distribution on different platforms (PCs, smartphones, tablets), and to generate visualizations and infographics. Most importantly, DDJ techniques allow the journalist to find news-value rich information in databases with thousands or millions of records, hardly manageable without the help of computers. Those techniques also facilitate the comparison between different databases for the production of new knowledge about society, creating mash-ups, for example, or reporting the results as text or audio-visual pieces.

According to Trasel [20], DDJ can be defined as "the application of computing and social sciences methods in the collection, processing, interpretation and presentation of data, with the aim of expanding the role of the press as advocate of public interest". From the ideas of authors such as Bradshaw [21] and Silver [22], one can synthesize the news production process from databases in three phases: collection, analysis and reporting. The collection phase includes collecting and standardization, or cleaning, of databases, that may be available in open government data portals or can be created by the reporter from original documents, surveys or through data capture using APIs, in the case of social networks like Twitter and Facebook. The analysis stage involves the processing of the collected data in order to identify newsworthy information. This phase also includes the interpretation of results and statistical experiments. Finally, the communication phase is the one in which the journalist produces texts, videos, audios and infographics based on the aforementioned interpretation of results.

Data visualizations, which are the main focus of this paper, are perhaps the product most valued by the contemporary press. Newspapers such as The New York Times and The Guardian have been gathering interdisciplinary teams of journalists, computer scientists and designers, in order to increase the production of infographics, data visualizations and animations, since they identify public demand for these materials. Furthermore, data visualizations are considered to be a way of communicating complex relationships between data more efficiently. It is a language and style choice that, from the perspective of journalists, facilitates their audience's understanding of issues subject to mathematical expression. Since the existence of well-informed citizens is an essential element in the maintenance of democratic societies, it becomes important to study the effectiveness of different types of data visualizations for communication. Therefore, this research aims to investigate the advantages and disadvantages of different kinds of data visualization in an applied context.

III. Related Work

Some previous works focused on analyzing whether the opinion expressed through Twitter can help to identify and track online sentiment. Some describe the mining process of texts and machine learning strategies to improve the processing of the collected content and results. Others argue how visualization techniques can help in data comprehension data and in its impact in various areas.

In Rodrigues Barbosa et al [23], the authors analyzed the hashtag's effectiveness as a resource for sentiment analysis expressed on Twitter. The results supported their hypothesis that hashtags may facilitate the detection and automatic tracking of online population sentiment about different events. Other authors such as Cunha et al [24] and Romero et al [25] also used hashtags supporting analysis of feelings on Twitter, with satisfactory results. All these works are similar to our considering the data collecting from social media and their processing so that we can characterize the sentiment (or opinion) of users about some subject. And in our case, tweets were hand-classified like negative, neutral or positive, based on the sentiment of the people in relation to a subject.

Pak et al [26] focused on using Twitter for the task of sentiment analysis. They show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They collected a corpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a

[3]The semifinal match, held on July 8th 2014, ended with an unseen score in world cups: Germany 7 vs. 1 Brazil.

set of objective texts (no sentiments). They build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document, with results efficient and satisfactory. Although they worked with English, the authors affirm the proposed technique can be used with any other language. Our work differs from this by using tweet text message only in Portuguese for sentiment analysis and by focusing mainly on the visualizations generated from Twitter's data about the 2014 World Cup, with the goal to understand how data visualization techniques can help data analysis.

Diakopoulos et al [27] looked at the first U.S. presidential debate in 2008, in conjunction with aggregated ratings of message sentiment from Twitter. They developed an analytical methodology and visual representations to help a journalist or public affairs person better understand the temporal dynamics of sentiment expressed in social media, in reaction to the debate video. This work is similar to ours in order to help journalists to meet their data needs and information. But we focused on visualizations techniques that can support the detection of opinions and new informations in social networks.

Khatua et al [28] explored the context of 2014 Indian General Election to test the predictive power of Twitter in a large and politically diversified country. The authors have analyzed roughly 0.4 million tweets and they observed that tweet volume as well as sentiment analysis can predict election results. They proposed a kind of template for data collection and cleaning and they found that sentiment scores can predict changes in vote share but, they emphasize the relevance of contextual understanding for efficient data collection and analysis. This work is similar to our considering the focus: sentiment analysis on Twitter. However, they focused on statistical analysis to build a model for the prediction of election result and we focused in the development of interactive visualizations techniques to present these data.

Satyanarayan et al [29] present a data visualization tool for journalistic storytelling. Informed by interviews with journalists, they introduced a model of storytelling abstractions. They evaluate the model through example applications and user studies with journalists. Study participants found the model to be a valuable prototyping tool that can empower journalists in the creation of interactive narratives. This work is similar to our considering the target public: journalists and their needs about interactive data visualization. In our case, the focus groups helped us to understand the needs of journalists in visualizing social media data. This is an essential requirement if we want, in near future, to develop a tool model that meets their needs.

Tabary et al [30] analyze occurrences of data journalism between 2011 and 2013 in Canada's Quebec province, discussing its actors, data access conditions, professional practices and the required computer skills. It traces the growing credibility of using quantitative and statistical evidences as sources of news stories, emphasizing the necessity of adequate training of the journalists to handle such production properly. Felle [31] debates data-driven journalism as an innovative method of investigating and presenting news stories relevant to the people on a larger scale. Through interviews with 26 journalists in 17 countries, the text also highlights a growing divide between the news teams that developed such data collection, interpretation and visualization skills and the ones that have not. Seen as an elitist tool, data-driven journalism is still distant from more popular news outlets, reducing the reach of its potential benefits.

All these works address related subjects to our interest and they were fundamental to understanding and learning about data collection processes from social networks, processing and presentation this data. We continue to further study all these issues with a focus on designing and developing data visualizations increasingly effective.

IV. Method

This is a descriptive research [32] focused on constructing a description about the opinions of journalists (either professionals or students) about some visualization techniques and how much these can help in their jobs.

The FIFA World Cup, held from June to July was one of the main discussion topics in Brazil in 2014 and reached high popularity on Twitter and other social networks. Anticipating the amount of social data related to this event motivated us to select them as a theme to the analysis. In order to do the analysis, we decided to start with a case study about the match "Germany 7 x 1 Brazil". For this, our work has been split into four steps: *i)* collect data from Twitter; *ii)* cleaning and annotation; *iii)* data visualizations; and *iv)* focus group sessions and analysis.

A. Data Collection, Cleaning and Annotation

In this study, we used a data set formed by collected messages from Twitter during the 2014 FIFA World Cup. This data set was called "WordCupBrazil2014" and contains 851,292 tweets in various languages (Portuguese, English and Spanish). These tweets were collected using the Twitter4J, based on Twitter Rest API, in the period between May 30th to July 13th, 2014. In order to collect tweets just referring to the World Cup, we defined a set of keywords that included the words "copa" (cup), "vencedor" (winner), "turistas" (tourists), "hexa" (informal abbreviation of six-time winner), etc. For each tweet, there is the following information: tweet id (tweet identification number), message (raw text), number of retweets (citation or reposting of a message), keyword (word used as filter during data collection process), timestamp (in BRST[4]), user id (user identification number who posted the message), hashtags, links and location (if available). This corpus was stored in a MySQL database.

[4]Brazilian Standard Time (GMT-0300).

Due to the result repercussion from the "Germany vs. Brazil" match (July 8th), we focused our study on messages posted during this match. To do so, we extracted a subset from the WorldCupBrazil2014 to form the "7x1-PT" corpus. This corpus contains 2,728 tweets in Portuguese, 35,024 tokens and 4,925 types. It is structured as follows: tweet id, timestamp, message (with preprocessed text) and polarity (manually annotated). Initially, we preprocessed the messages to facilitate the annotation of polarity. We removed hashtags and URL links from the message body. This automatic cleaning was not very simple, because there were messages where the hashtags were doing part of the sentence structure. These hashtags had syntactic functions of subject or object, as in "Na #Torcida pelo #Brasil" (#Cheering for #Brazil). In order to resolve this problem, in most cases, we just removed the character #: "Na Torcida pelo Brasil". This step was manually refined. Other difficult case was the ad hoc abbreviations commonly made by the users. A regular abbreviation pattern is the omission of vowels, as q that means "que" (that). In order to facilitate the transformation of abbreviations, we have created a list from Wikipedia, as in [33], which related an abbreviation to corresponding word. In this case, we had problems when messages did not separate the words by some delimiter as white space or comma, as in "oq" meaning "o que" (what).

The polarity annotation was based on the sentiment of the people in relation to Brazil national soccer teams performance. We manually annotated each tweet as negative, neutral or positive. We considered as positive those tweets that praised the Brazilian team or that expressed encouraging messages. We considered as negative those tweets that criticized the Brazilian teams performance or that expressed pessimistic messages. The remaining tweets were considered neutral. The assignment of sentiment labels to messages is a very subjective task and it is often difficult because it depends on the message context and personal interpretation of the human annotator. For this reason, this process was carried out by two human annotators. Although the initial observed agreement between annotators had been 0.53 [34], the second annotator had as its main function to discuss and to review the label defined by the first annotator, especially when the polarity of the message was not clear. For example, the tweet "A Copa das copas" (The Cup of the cups) at the beginning of the game, had positive polarity. However, from the fifth goal of Germany on, this message turned to be posted ironically. At each German goal, we perceived an increase in posts with humor, satirizing the performance of the Brazilian team. It is worth mentioning that irony and satire have been noted with neutral polarity. Besides, at the beginning of the game, mostly, there were tweets from different domains and purposes. Some tweets were too short, formed only by verbs such as "Começou!" ("Began!"). Others tweets were referring to political, advertising or others news. All these tweets were classified as neutral.

Table I
POLARITY DISTRIBUTION OF THE 7X1 CORPUS.

Polarity	# Tweets (%)
Negative	800 (29%)
Neutral	1,771 (65%)
Positive	157 (06%)

Table I displays the distribution of tweets with respect to polarity.

B. Data Visualization: understanding changing contexts

As mentioned before, visualization techniques are able to bring more information as well as to assist decision making in various levels and contexts. In the context of data-driven journalistic practices and products, they are used to translate huge amounts of data into accessible visual solutions, illuminating the understanding of complex issues. As the presented visualizations also suggest, and considering the polarity annotation, they are able to reflect shifts in mood and opinion, as we complement the visualizations with a few contextual notes.

Some of the hashtags most used by online Brazilians about the World Cup before the event took place were #vaitercopa (#therewillbeworldcup), #vaitercopasim (#therewillbeworldcupyes) and #naovaitercopa (#therewillnotbeworldcup). They were often used with a political bias to emphasize the opinion of the user about the preparations for the sport event as praises and critics of public spending on stadiums, delays on urban and mobility works, and so on, voiced their opinions online. Sensing the polemic aspect of the online debate, late in 2013 the Brazilian government suggested a different hashtag to be adopted, #copadascopas (#worldcupofworldcups), which became part of public speeches by president Dilma Rousseff and was echoed in the press and other official channels, but did not reach widespread popularity.

However, the divisive hashtags continued to be used during the sport event and therefore spiked our research interest in understanding this use. During the data collection process we created some simple bar and line charts related to the visualization of hashtags, as shown in figure 1. This bar chart presents the percentage of the use of the three hashtags during the whole day of July 8th, when the semifinal game took place.

To add some context to the bar chart, we highlight that the afternoon pre-game sees an increase in the favorable hashtags, right up to the beginning of the game at 17h. As the German team scored, ending the first half with 5x0 and finishing the second half with 7x1 around 19h the percentages of hashtags change accordingly, with #naovaitercopa (#therewillbenocup) growing in proportion.

In addition, we were interested to know what were the ten most used hashtags by online Brazilians on Twitter during the 8th of July. This is shown in the line chart in figure 2.

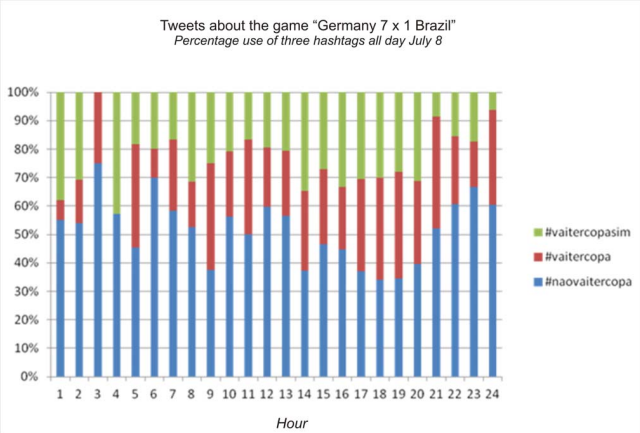

Figure 1. Percentage of tweets with the three hashtags that suggest whether there will or will not be World Cup.

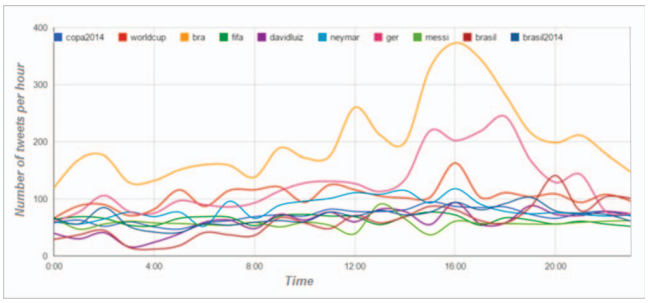

Figure 2. Most frequent hashtags that appeared on July 8th.

Both visualizations (figures 1 and 2) show only the frequency of tweets containing such hashtags. Our interest was to verify whether it would be possible to extract relevant information from it. Again, the shifts in the curves suggest possible interpretations as the hashtag #Bra peaks in the pre-game moments and is seen falling as the #Ger goals sealed the opponent’s victory in the semifinal. Accordingly, #Ger also rises towards the end of the game. Figure 2, however, also points to an unusual activity of #Bra around noon, what can suggest a synergy between the schedules of sports news broadcasts on TV and Radio and Twitter usage.

The next visualization created in the process show data from the processed results. Figure 3 shows sentiments on Twitter about the attacking players of the Brazilian team, classified as positive, neutral and negative. The X-axis contains the score of the game changed, including also the break between the first and second half.

Besides the use of the popular static bar and line charts showed in figures 1, 2 and 3, we also developed a prototype for the interactive visualization of sentiment analysis obtained with the polarity annotation. First, it is important to say that we use the view of Pang et al [35] to conceptualize the terms “opinion mining” and “sentiment analysis”. The authors pointed out that both terms denote the same field

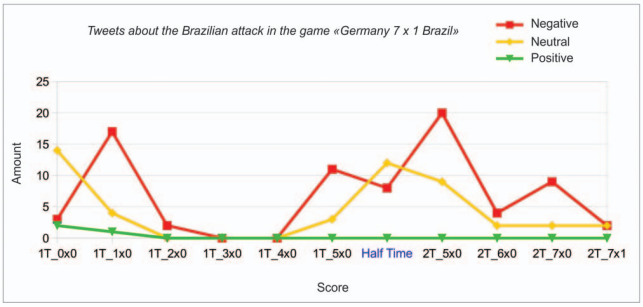

Figure 3. Tweets about the attacking players of the Brazilian team classified as positive, neutral and negative.

of study (which itself can be considered a sub-area of subjectivity analysis). The terms have multiple definitions that covering the subjective textual analysis of social media source.

A glance reveals there was little optimism present on the Brazilian Twitter crowd, possibly reflecting the absence of the teams’ star player Neymar Jr. injured in the previous match. Other spikes in the graphic mark the first goal of Germany and the beginning of the second half of the game when the negative comments suggest the understanding that a rags-to-riches victory was impossible. Curiously, the other goals scored by Germany in the first and second half provoked less reactions as measured by our research.

Figures 4 and 5 shows the two interactive visualizations of sentiment analysis about the semifinal game already implemented in the prototype, in which tweets were classified as positive, neutral or negative. These prototypes were developed using the D3.js [36]. In figure 4, the visualization corresponds to the average of the tweets classified by the three kinds of sentiments during the game. The user can interact on the timeline that appears below the graphic, zooming the period of time to be shown in the chart through the selection of a part of the timeline. In figure 5 the visualization shows the amount of tweets by kind of sentiment during the game, and the interaction also occurs through the selection of a period of time in the timeline to do a zoom in or zoom out.

C. Focus Group Sessions and Analysis: inviting outside looks

According to Tremblay [37] focus groups are a relevant approach for refining and evaluating design artifacts for several reasons, such as its ability to allow for the emergence of ideas or opinions that are not usually revealed in individual interviews, and the possibility to collect rich data from subjects. Focus groups offer design researchers a technique employed at various moments of the process, to collect user needs, for feedback on concept sketches or prototypes, or to allow participants to generate new ideas. Focus groups can also be used for final concepts refinement.

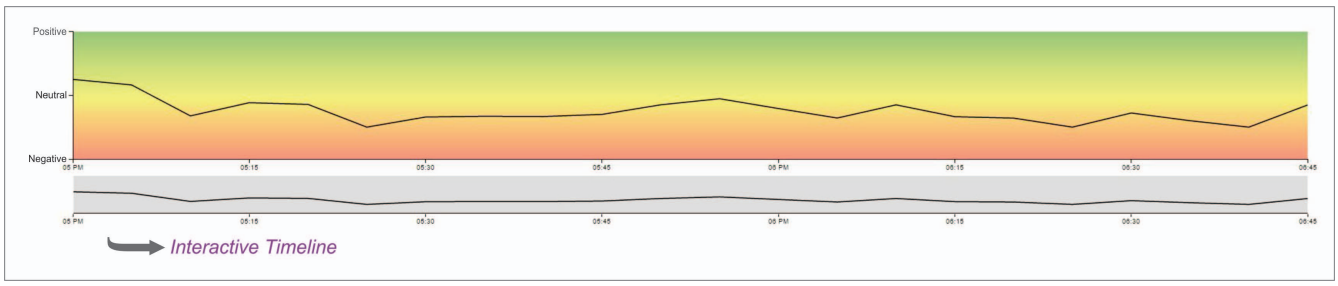

Figure 4. Tweets from the "Germany 7 x 1 Brazil" match, classified as positive, neutral and negative, were visualized through the implemented prototype as an average.

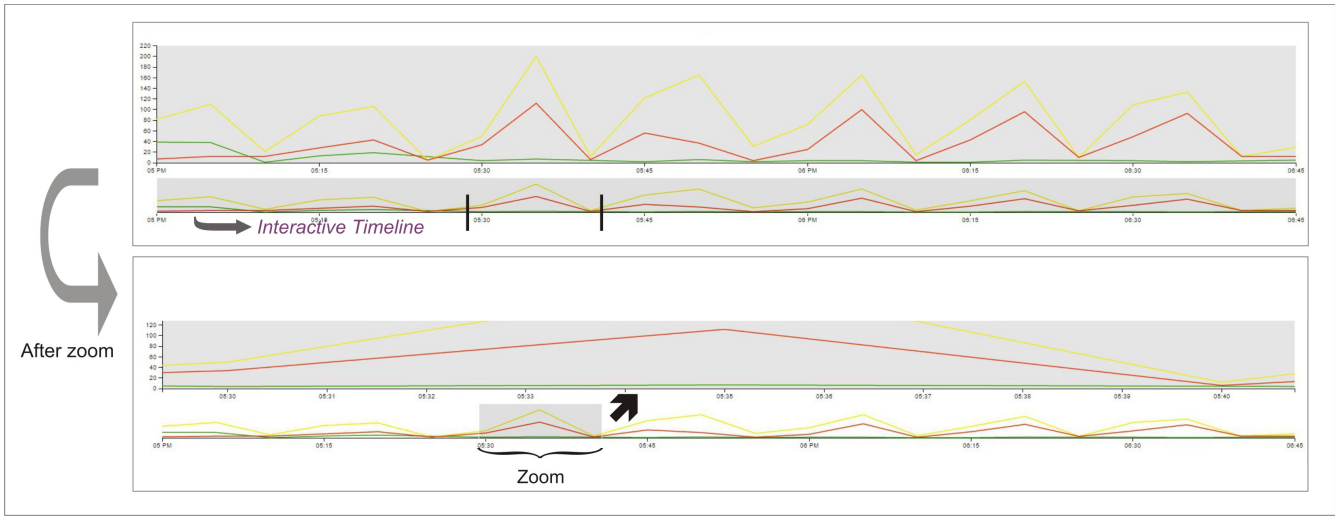

Figure 5. Tweets from the "Germany 7 x 1 Brazil" match, classified as positive, neutral and negative, and the result of the zoom done through the selection of a small period of time in the interactive timeline.

Lazar et al [32] highlights that the implementation of two or more focus groups will increase chances of success, because relying on a single focus group session is discouraged, as any single group could be unresponsive or unrepresentative. In this work, two focus group sessions were conducted to understand and explore human-data interaction, specifically between journalists (either professionals or students) and the provided visualization techniques. We intended to get their opinions about the created visualizations with data collected on Twitter from the Brazil vs. Germany game in the 2014 World Cup. The main question that guided the focus group sessions was: *what information can be obtained from the presented visualizations?* The intended outcome of the focus group was to synthesize the key elements of a discussion revolving around the existing elements presented in a visualization (either interactive or not).

With this aim, the focus group sessions were conducted in different days in a private room in the University, lasting about 2 hours, and four visualizations techniques were presented to the participants by a moderator. In the first focus group session, 4 individuals (female) participated in the study, all professional journalists as well as graduate students, with a mean age of 27 (between 26 and 30). In the second session, 5 individuals (2 female and 3 male) participated in the study, all undergraduate students in Journalism with active participation on an experimental convergent-media newsroom, with a mean age of 19 (between 18 and 22).

All participants have accounts on Twitter and often use this medium as source of information, but only 4 of them use Twitter to disseminate information. They all said that they did not use any application for data analysis in social networks. When asked about the benefits of using social networks in journalistic activities, the participants pointed out advantages such as: supplemental information, source of the news lead (agenda), search by facts and cases, quickly and easily disseminate information and get supplies, monitor impact of facts and information, among others.

During the focus group sessions, following a short explanation of each visualization technique, the mediator encouraged the group's participants to talk about their insights, without a set time limitation, and to openly discuss their

perceptions and experiences as well as the pertinent qualities of the visualizations with each other. For the two interactive visualizations, the mediator invited participants to go to the computer and interact with them.

The focus group sessions were recorded using two video cameras (for redundancy) and an audio recording device. The transcriptions of these recordings and annotations about discussion topics are the main component of the analysis presented in the next section.

V. Results

The results presented in this section came from the analysis of the focus group sessions discussions and were organized under two major aspects: journalistic criteria and visualization techniques.

To allow a better understanding of the discussion, the visualization techniques were nominated as:

- Graphic A: percentage of use of the three hashtags during all July 8th (figure 1).
- Graphic B: ten hashtags most used by online Brazilians during July 8th (figure 2).
- Graphic C: tweets about the attacking players of the Brazilian team (figure 3).
- Graphic D: sentiment analysis of tweets from the "Germany 7 x 1 Brazil" match (figure 4 and 5).

From the viewpoint of communication, the results of the focus group sessions suggest that data visualization, in general, provides the ability to identify trends, patterns and anomalies in datasets which would not be visible to the naked eye. As such, they have an intrinsic value to complement evidence collected by journalists reporting on a subject.

Both focus group sessions converged on key points detailed below, with one notable exception: student journalists tended to be more critical of the visualizations, often suggesting style and color layout changes and possible uses with other areas of communication as advertising. We attribute this difference to the interdisciplinary ambience of the work in the communications school, but also to the speed and objectivity that professional journalism demands from its staff, something that might eclipse speculative solutions easily.

The graphic A, which charted three hashtags usually understood as suportive or critical of the World Cup, was understood as relevant by both groups of participants. The student journalists criticized the color scheme, where the positive hashtag "#vaitercopa" (#therewillbeworldcup) was plotted as red, a hue usually associated with some interruption (as in a traffic light), a sign of caution (as in the STOP street sign) and not with a upbeat emotion. As a counterpoint, both groups questioned the multiple possible readings of the hashtags, where the same text can be used with an ironic or satirical intention with the addition of other words or tags, something the dataset preparation and the following visualization did not contemplate and remains to be implemented.

The regular time intervals of the columns of the hashtags in the graphic A also deserved a mild criticism by both groups. A recurrent suggestion for this visualization was the superimposition of a timeline indicating the key moments of the day related to the game, making it easy to draw conclusions from the graphic. A relevant question related to the visualization asked why it did show the bars on hourly intervals and did not reflect occurrences between them. Although the researchers agree that such precision would be desirable, a substantial shift in opinion on Twitter possibly would ripple for hours and thus could be represented on the next bar.

In the case of graphic B, for example, both focus group sessions' members identified in the upslope and peak mentions of the keyword BRA the rising expectation of the Brazilian fans before the game. Although a user watching the data stream on Twitter through his timeline would be able to see a lot of references to BRA, depending on the preferences and interests of the profiles that he follows, he would hardly be able to determine the differences of proportion between mentions of this and other terms. The presentation of these proportions as lines on a chart, however, allowed the evaluation of how the term BRA was adopted by fans on the day of the match between the Brazilian and German teams.

The data visualizations also enables the identification of counter intuitive patterns. For the same graphic B, participants in both groups wondered, for example, why the name of the player Messi, of the Argentinian national team, was among the most cited keywords on a day when there was no match of his team. It is not in the scope of this research to investigate the reasons for this, but the surprise of one of the participants suggests the identification of an unexpected phenomenon, which could, in the context of journalism, generate a pitch for a news article[5].

Regarding criticisms to the graphic B, participants of the second focus group session suggested the visualization would be more readable if it showed less lines, choosing only the ones directly related to the teams participating in the match. The researchers consider this to be a relevant point, but such a choice would obscure the understanding of uncommon patterns as shown by the frequent mentions to the

[5]Possible explanations to this occurrence are varied and involve the relation of twitter with other media. First of all, on the same day of the Brazil vs. Germany game, Argentina was en route to São Paulo to a match with Holland the following day, a fact that generated media attention and subsequent twitter chatter. The television broadcasts of sports news around 13h did report on this and helped spread the online dialogue. As a second factor, the Brazilian newsmagazine with the highest circulation was running a digital vote to elect in the fans views who was the World Cups best star player. Third, and perhaps more important, spontaneous tweeting by Brazilians explored the long standing soccer rivalry with Argentina, wishing for a face-to-face duel between the best players (Neymar Jr. and Messi) either on the final match or in the match for the third place.

player Lionel Messi as discussed above. A balance between clarity and observation could be attained if the journalist could choose the number of lines to be viewed, elaborating its arguments from such selection. Other points raised were the possible addition of an interactive zoom features on the graphic B, once more the inclusion of a timeline indicating the key events of the day and a more evident mention that the tweets analyzed were from a portuguese-only dataset.

The graphic C which represented the sentiment analysis regarding the Brazilian team's attack players, was praised for its clarity by the participants, who related the shifts in the positive and negative lines to recollections of the widespread mood in the country on the day of the match. Participants of the second focus group session manifested the curiosity to keep following the sentiment shifts beyond the end of the game, observing how the nation's Twitter users dealt with the despicable score.

However, the participants also showed relevant criticism to the graphic C. A participant from the first group commented: *"The neutral and positive are well within the expected, but what those two major negative peaks are, I do not know just from looking at the chart"*. In the participants interpretation, the display showed abnormalities which could be investigated by diverse strategies: the ability to stream a video recording of the match, to mouse over the lines and see the changes in the word frequency during the match and a content analysis of the tweets considered negative, for example, could qualify an in-depth understanding of the changing sentiment.

Another criticism made to the graphic C was constant space used between points in the X-axis, which suggested that the goals happened in regular intervals of time, what was very different from fact. As the FIFA report on the match shows, four out of the seven goals happened within six minutes, between the twenty-third and twenty-ninth minutes of the first half[6]. The lines reflect the surprise, falling drastically in the second and third German scores, with a negative spike coming right after. Even so, the researchers consider this to be a valid criticism, as a clear proportion between space and time representation could inform a more realistic interpretation of the data and that day's event.

In general, the participants considered the interactive graphic D (figures 4 and 5) as the most friendly and interesting among all presented. The graphic D comprised two interactive views of sentiment analysis on Twitter about the referred match in the World Cup. The preview option on separate lines for each sentiment was praised by the participants of both focus group session, as it made it possible to identify the peaks of negativity or positivity and elaborate arguments from such visualization.

About the visualization with the average of sentiments, focus group sessions' participants also discussed the fact that ironic sentiments during the game occurred, which may have influenced the average neutral sentiments, since the irony would probably not have been considered a positive or negative sentiment.

In fact, one of the participants stated that, if there were displayed first separate lines and then the option to visualize the average, one could not imagine that the result of the average sentiment was neutral, but would think the general attitude during the game was negative. This discrepancy between the perceptions generated by the two visualization options from the same data points to the importance of taking into account the abilities and limitations of the human visual system for the identification of standards [10] when creating data visualizations. In this sense, interactivity is a productive strategy to facilitate understanding of large amounts of data, because it allows the user to observe and compare different types of patterns and thus observe the same database under different perspectives.

As a suggestion for improvements to the visualizations presented, the participants in both groups suggested greater possibility of interaction as *"when you hover the mouse over the chart line, show the exact number of tweets"*, *"superimpose the game timeline with a video recording of the game, to facilitate understanding of the peaks of tweets on the chart and relate them to the time of the game in which that peak occurred"*. In addition, another suggestion was *"to create a connection from the timeline with news of the moment, for a better understanding of the behavior of users over time"* and the possible connections between more traditional media and Twitter.

In this sense, the focus group sessions emphasized that visualization techniques are a relevant tool for identifying deviant patterns. For journalists, these patterns are important, as they generally indicate events endowed of news value events whose attributes make them prone to be included in the news. To paraphrase the traditional formula: visualizations can often show the moments in which man bites dog, that is, when expectations about the progress of daily life are reversed. Based on the identification of these phenomena, reporters may seek human sources and documents that explain its context, in order to inform the audience about that event. If journalists depended on the mere monitoring of timelines, these patterns could remain buried under the gibberish of updates and, in practice, remain invisible. Even before they are news products, then, data visualizations are tools to assist in the construction of news.

VI. Conclusions and Future Works

In this paper we presented a study about how data visualization techniques can support media studies and journalists in data comprehension and in extraction of new information. We collected and processed tweets about the 2014 World Cup and we created visualizations about the users' sentiment

[6]2014 FIFA World Cup Brazil Match report: Brazil - Germany 1 : 7. Available online on: http://goo.gl/AN2Rv7

in "Brazil vs. Germany" game, and about hashtags most used by online Brazilians on Twitter. Through the visualizations it was possible to see how these sentiments were expressed and propagated in Brazil. Furthermore, we realize that hashtags do carry information that can complement the data analysis about a subject. Besides that, although the visualizations were not made in real time, it became clear that their use can help spot emerging subjects and discussion points, setting the agenda of news stories on relevant yet unusual themes.

After data preprocess, visualizations were created and presented to journalists in two focus group sessions. The analysis of their comments revealed some relevant points. It is possible to accent first the consistent favorable position regarding the use of visualization techniques as a valuable resource in journalism. Second, the need that a visualization must be associated with contextual clues or explanations which help one make sense of it. Third, data visualizations can also operate as conversation starters. Fourth, there was a preference for interactive visualizations. Thus, we notice that even simple visualizations techniques as bar chart can help journalists to have insights, understand the data and provide distinguished information to readers.

Analyzing comments and reactions of the participants, we realize that data visualizations are relevant tools to support the construction of news stories and in depth understanding of events, and the journalists pointed out features that could improve the visualizations presented such as the increase of interactive visualization techniques.

We know that simply having these analysis is far from our eventual goal of establishing one or more models of interaction in data visualization. However, we believe that these analysis are an initial step toward this direction. We believe that understanding needs of users (data analysts) related to interaction techniques would provide a framework approaching to providing a good model of interactive visualizations. We are currently conducting further focus groups and interview to collect feedback on the visualizations discussed in this text. The next group to be considered are journalists from the five regions of Brazil, a step we understand as coming full circle on our main potential audiences: student journalists, journalists with an academic presence and professional journalists from national newsrooms. Also, in order to qualify the gathering process and the visualizations produced, as next steps, we envisage activities such as:

- Replace manual data preprocess works by automated techniques.
- Include emoticons in the identification of sentiments and still use the repetition of vowels to denote the intensity of these sentiments.
- Relate the visualizations with the news events of the day and scheduling of mainstream media.
- Create different versions of the same data visualization to try and find which worked best at getting the message across, as well as improve and develop further interactive visualizations.

Thus, as future work, we intend to deepen discussion and evaluation of interaction techniques to lay an initial foundation toward a deeper understanding about interactive visualizations, especially related to users' needs, and to develop new prototypes and tests. Analyze the content posted on Twitter is complex, subject to ambiguity and errors, and raises several technological challenges. How to treat sarcasm and irony - frequent issues in our country - and how to present it in new interactive visualizations is one of the biggest challenges in this context.

ACKNOWLEDGEMENT

This work was partially supported by PUCRS (Editais 07/2014 e 07/2015 - Programa de Apoio a Integração entre Áreas/PRAIAS). The authors acknowledge the collaboration of Prof. Rodrigo Barros and of the students Cinara O. Padilha, Eduardo P. L. Hoefel, Jocines D. da Silveira, Lorenzo P. Leuck and Juvenal N. S. dos Santos Jnior.

REFERENCES

[1] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, Inc, 2013.

[2] P. Lévy and R. Bononno, *Becoming virtual: reality in the digital age*. Da Capo Press, Incorporated, 1998.

[3] A. Hanna, C. Wells, P. Maurer, L. Friedland, D. Shah, and J. Matthes, "Partisan alignments and political polarization online: A computational approach to understanding the french and us presidential elections," in *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, ser. PLEAD '13, 2013, pp. 15–22.

[4] T. N. Smyth and M. L. Best, "Tweet to trust: Social media and elections in west africa," in *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers - Volume 1*, ser. ICTD '13, 2013, pp. 133–141.

[5] R. Cohen and D. Ruths, "Classifying political orientation on twitter: Its not easy!" in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Publications, 2013, pp. 91–99.

[6] M. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, Oct 2011, pp. 192–199.

[7] F. A. Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, 2012, pp. 387–390.

[8] A. Sahami Shirazi, M. Rohs, R. Schleicher, S. Kratz, A. Müller, and A. Schmidt, "Real-time nonverbal opinion sharing through mobile phones during sports events," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, 2011, pp. 307–310.

[9] J. Chen, A. Cypher, C. Drews, and J. Nichols, "Crowde: Filtering tweets for direct customer engagements," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Publications, 2013, pp. 51–60.

[10] 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.

[11] M. Elias and A. Bezerianos, "Exploration views: understanding dashboard creation and customization for visualization novices," in *Human-Computer Interaction–INTERACT 2011*. Springer Berlin Heidelberg, 2011, pp. 274–291.

[12] M. Elias, M.-A. Aufaure, and A. Bezerianos, "Storytelling in visual analytics tools for business intelligence," in *Human-Computer Interaction–INTERACT 2013*. Springer Berlin Heidelberg, 2013, pp. 280–297.

[13] J. Heer, F. B. Viégas, and M. Wattenberg, "Voyagers and voyeurs: Supporting asynchronous collaborative information visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07, 2007, pp. 1029–1038.

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

[15] S. Parasie and E. Dagiral, "Des journalistes enfin libérés de leurs sources? promesse et réalité du" journalisme de données"," *Sur le journalisme About journalism Sobre jornalismo*, vol. 2, no. 1, pp. pp–52, 2013.

[16] C. W. Anderson, "Notes towards an analysis of computational journalism," *HIIG Discussion Paper Series*, vol. 2012-1, 2011.

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

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

[19] 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.

[20] M. R. Träsel, "Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no brasil," Ph.D. dissertation, Pontifícia Universidade Católica do Rio Grande do Sul, 2014.

[21] P. Bradshaw, "The inverted pyramid of data journalism," *Hg. v. Online Journalism Blog*, vol. 7, 2011.

[22] N. Silver, "What the Fox Knows," FiveThirtyEight http://fivethirtyeight.com/features/what-the-fox-knows/, 2014.

[23] G. A. Rodrigues Barbosa, I. S. Silva, M. Zaki, W. Meira, Jr., R. O. Prates, and A. Veloso, "Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment," in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '12, 2012, pp. 2621–2626.

[24] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto, "Analyzing the dynamic evolution of hashtags on twitter: A language-based approach," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11, 2011, pp. 58–65.

[25] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11, 2011, pp. 695–704.

[26] 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.

[27] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10, 2010, pp. 1195–1198.

[28] A. Khatua, A. Khatua, K. Ghosh, and N. Chaki, "Can #twitter_trends predict election results? evidence from 2014 indian general election," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, Jan 2015, pp. 1676–1685.

[29] 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.

[30] C. Tabary, A.-M. Provost, and A. Trottier, "Data journalism's actors, practices and skills: A case study from quebec," *Journalism*, p. 1464884915593245, 2015.

[31] T. Felle, "Digital watchdogs? data reporting and the news media's traditional 'fourth estate' function," *Journalism*, p. 1464884915593246, 2015.

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

[33] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11, 2011, pp. 30–38.

[34] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[35] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[36] S. Murray, *Interactive data visualization for the Web*. O'Reilly Media, Inc, 2013.

[37] M. C. Tremblay, A. R. Hevner, and D. J. Berndt, "Focus groups for artifact refinement and evaluation in design research," *Communications of the Association for Information Systems*, vol. 26, pp. 599–618, 2010.