Sentiment Analysis on Brazilian News Broadcast Data

Alexandre Martins da Cunha[1], Isabela Santos[1], Daniel Pedrosa[1], Francis F. Steen[2], Mark Turner[3], Maira Avelar[4], Lilian Ferrari[5], Gustavo Paiva Guedes[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

[2]University of California, Los Angeles - Los Angeles, CA 90095, USA

[3]Case Western Reserve University - 10900 Euclid Ave, Cleveland, OH 44106, USA

[4]UESB - Estrada Itapetinga/Itambé - s/n, Itapetinga - BA, 45700-000

[5]UFRJ - Universidade Federal do Rio de Janeiro
Av. Horácio de Macedo, 2151 - CEP 21941-917 - Rio de Janeiro - RJ - Brasil

alexandre.cunha@eic.cefet-rj.br, bela.rsantos@gmail.com,
daniel.souzapedroza@gmail.com,steen@comm.ucla.edu,
mark.turner@case.edu, mairavelar@uesb.edu.br,
lilianferrari@uol.com.br, gustavo.guedes@cefet-rj.br

Abstract. *This work aims to compare the occurrence of negative emotion words in Brazilian broadcast news JN and JR, and also analyzes Twitter posts related to them. We use the Brazilian Portuguese version of LIWC dictionary, which is a Sentiment Analysis software. The results indicate that both JN and JR tend to use negative emotion words, but in JR this tendency is greater. Nevertheless, Twitter posts direct more criticisms towards JN than JR.*

1. Introduction

The world is constantly changing and the relationship between public and press is also evolving over time. Information and communication technology (ICT) allowed internet users to access, transmit and manipulate information. In fact, information propagation has been growing in the free network environment in ways that would be unimaginable several years ago [Ma et al., 2013]. For example, online social networks (OSNs) provide a new way of spreading information which is far beyond "word-of-mouth" [Ma et al., 2008]. Recent studies, however, indicate that television still remains the main communication medium [Gamonar, 2015]. In Brazil, 95% percent of the people watch television regularly and 74% watch it every day [Vizeu, 2016], and telejournalism is the most important communication medium in the country [Vizeu, 2016].

Given that it has been attested that criminal-and violence-related news can increase the audience of a news channel [Junqueira et al., 2013], the preference for news with negative focus in Brazilian telejournalism comes as no surprise [Vaz and Medeiros, 2014]. In order to evaluate the exhibition of negative news in Brazilian telejournalism, this work analyzes Closed Captions (CC) from Brazilian news broadcasts. Since Sentiment Analysis (SA) has become a hot topic in recent years [Pang et al., 2008], we analyzed the CC with a SA software named Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al.,

2001]. Finally, we applied *Cohen's effect size* to analyze psycolinguistic differences of word occurrences in the CC.

The data set analyzed in this work consists of CC from two of the most-watched news program in Brazilian television (i.e., Jornal Nacional (JN) and Jornal da Record (JR)) extracted from 05/22/2017 and 12/06/2017. The first one is exhibited by rede Globo® and the latter is exhibited by Rede Record®. Closed Captions were captured by the Brazilian Red Hen Capture Station[1], with the facilities of the Distributed Little Red Hen Lab.

Experimental results indicate that there are many negative news reports in both JN and JR. However, JR has more negative words (indicating anger, sadness, death) than JN. After obtaining these findings, we analyzed Twitter posts related to JN and JR, based on the assumption that there would be several criticisms of the negative content of Brazilian broadcast news [Ribeiro, 2015]. To investigate the extent of negative feelings about JN and JR, we checked Twitter posts for a range of negative keywords referring to JN and JR. For example, the negative word "desgraça" (disgrace) appeared in some collected twitter posts, as in "JN só mostra desgraça" and "JR só mostra desgraça", which roughly mean that news broadcasted in these channels report on bad things. Preliminary results indicate that there are more criticisms of JN than JR.

The remainder of this paper is organized as follows. Section 2 presents related work on Brazilian broadcast news analysis. Section 3 outlines main LIWC characteristics. Section 4 discusses data set acquisition. Section 5 explains the methodology used to analyze the data. Section 6 summarizes our experiments. We conclude in Section 7.

2. Related Word

Given the nature of our work, we focus on publications directly related to news broadcasts. In Junqueira et al. [2013], the authors analyze the representation of urban violence in telejournalism. They analyze content and discourse in urban violence news, using two newsletters: *Bom dia Goiás*, from tv Anhanguera, a subsidiary of Rede Globo® and *Direto da redação*, from Record Goiás, a subsidiary of Record®. They conclude that both newspapers violate the principle of human dignity; however, the newsletter *Direto da redação* produces more serious violations of citizenship. The authors conclude that the main goal of *Direto da redação* is to transmit rapid, superficial and decontextualized information.

Vaz and Medeiros [2014] focus on the news production process, which involves events selection. The aim is to investigate the negative aspect of facts converted into news. Their work performs an analysis in news extracted from *UOL* web site, *Folha de S.Paulo* newspaper and the *Jornal Nacional* broadcast news. The work concludes that the abusive use of negative aspects of facts should be questioned.

In Almeida [2017], the authors criticize the way telejournalism generally reports and discloses criminal trials. They emphasize that in some news broadcasts, journalists dramatize human pain in scenes of dead people, looking for a guilty party against whom society can turn. They conclude by pointing out that in many cases, telejournalists extrapolate their obligation to inform, and often exhibit only violence and death.

[1] http://www.redhenlab.org/

Our study also analyzes negative aspects in news. However, we focus on a direct comparison of CCs from two of the most watched news program in Brazilian television, exhibited by Rede Globo® and Record®. Our work also differs from those mentioned above by using a reputable sentiment analysis software. We also analyze twitter messages in order to evaluate audience's feelings about these two news program.

3. LIWC

LIWC (Linguistic Inquiry and Word Count) is a software that can obtain narrative, structural, emotional and intellectual elements from written texts [Pennebaker et al., 2001]. In the early days, LIWC was used to improve mental health treatment through analysis of patient narratives about negative experiences [Pennebaker et al., 2003]. In the course of time, new applications have been proposed, such as the transcription of daily narratives [Pennebaker et al., 2003].

LIWC encompasses a large collection of entries, in which each entry is associated with one or more categories. These categories are related to linguistic processes (e.g., pronoun, verb, article) and psychological processes (e.g., anxiety, negative emotion, swear words) [Pennebaker et al., 2003]. The Brazilian Portuguese version of LIWC is based on the 2007 LIWC English dictionary and has 127,149 entries divided into 64 categories [Balage Filho et al., 2013].

4. Data Set

In order to investigate the exhibition of negative news in Brazilian telejournalism, this work analyzes text from news broadcasts. To achieve this goal, we created a data set named `NewsBroadcast-PT-2017`, which comprises collected Closed Captions from JN and JR in the period between 05/22/2017 and 10/14/2017. The data were collected with the facilities of the Distributed Little RedHen Lab.

The data set comprehends CCs from 170 news programs. It contains 85 CCs from JN programs with duration between 44 and 60 minutes. On the other hand, it contains also 85 CCs from JR programs with duration between 45 and 55 minutes. It is important to mention that each CC from JN has, for the same date, the correspondent JR CC.

5. Methodology

This section describes the methodology adopted to explore the sentiment analysis in CCs. As described in Section 1, there is a preference for news with negative focus in Brazilian telejournalism, since news about crime and violence can increase the audience. Thus, we used the `NewsBroadcast-PT-2017` data set to analyze the use of the following LIWC categories: negation words (*negate*), negative emotion words (*negemo*) and positive emotion words (*posemo*). It is important to note that *negemo* is divided into three subcategories: anxiety words (*anx*), anger words (*anger*) and sadness words (*sad*).

In this scenario, we first created a frequency vector using the above mentioned categories for each CC, as illustrated in Figure 1. As an example, in this vector, the frequency of $negate$ words is 53.

Then, in order to investigate psycholinguistic differences between news from JN and JR, we conducted an effect size analysis using Cohen's d statistic [Rosnow and Rosenthal, 1996], as shown in Eq. 1. In this equation, i represents the LIWC category of

$negate$	$posemo$	$negemo$	anx	$anger$	sad
53	20	14	4	7	3

Figura 1. Frequency vector representing a CC with 6 dimensions.

examination ($0 \leq i < 6$), $\bar{X}^i_{JN}$ and $\bar{X}^i_{JR}$ simple averages of the ith component of the frequency vectors of JN and JR respectively. Likewise, SD^i_{JN} e SD^i_{JR} are the standard deviations of words for each category i.

$$d_i = \frac{\bar{X}^i_{JN} - \bar{X}^i_{JR}}{\sqrt{((SD^i_{JN})^2 + (SD^i_{JR})^2)/2}} \tag{1}$$

The intuition for the interpretation of this equation is described as follows. Positive values of d_i indicate that JN used more words in category i than those of JR. Negative values denote greater use by JR than JN. Cohen proposed the interpretation of $d = 0.20$ as small effects, $d = 0.50$ as medium effects and $d = 0.80$ as large effects [Cohen, 1988].

6. Results

In order to analyze the differences between CCs from JN and JR, Table 1 presents the results of the calculation of Cohen's effect size (Eq. 1) for each of the following LIWC categories: *negate*, *negemo*, *posemo*, *negemo*, *anx*, *anger* and *sad*. For these groups, mean values ($\bar{X}$) and standard deviation (SD) were also presented. Positive values of d indicate that CCs from JN presents more words than JR in the correspondent category. Analogously, negative values indicate that CCs from JR contains more words related to the correspondent category.

Tabela 1. Positive values indicate the JN used the category more than the JR. The size of the effect is represented by d, according to Eq. 1) $\bar{X}$ and SD represent the mean and standard deviation, respectively.

		JN		JR		
Category	Example	$\bar{X}$	SD	$\bar{X}$	SD	(d)
negate	not, never	0,86	0,30	0,62	0,27	0,82
posemo	love, best	3,36	0,39	3,07	0,35	0,80
negemo	afraid, cry	1,86	0,29	1,91	0,33	-0,18
anger	hate, raping	0,72	0,17	0,77	0,19	-0,27
sad	crying, sad	0,56	0,11	0,60	0,15	-0,27
death	kill, war	0,19	0,09	0,28	0,12	-0,85

It can be noted that JN has a tendency to use more negation words (*negative*) in news. On the other hand, it also shows that JN CCs present more words related to positive emotion (*posemo*). In contrast, JR presents more negative emotion words (*negemo*), anger words (*anger*, sadness words (*sad*) and death words (*death*). It is important to note that death-related words present a large effect, indicating that JR exhibits much more death-related words than JN.

We have selected some negative keywords to analyze negative feelings about JN and JR in Twitter posts. We combined these keywords with "Jornal Nacional", "Jornal da Record", and "só" (only). We also used "*" to achieve better results. Tweets were collected between January 2010 and February 2018. Table 2 shows results for queries "Jornal Nacional só * desgraça" and "Jornal da Record só * desgraça", which means roughly that these channels report only news with content related to "disgrace". Results show 156 twitter messages in the JN search and 46 in the JR search.

Table 3 shows the results combining the combination of four negative keywords with "Jornal Nacional" and "Jornal da Record". The keywords are: "notícia ruim" (bad news), "tragédia" (tragedy), "morte" (death) and "violência" (violency). Results shows 155 twitter messages in the JN search and 28 in the JR search.

Tabela 2. Twitter search query including keyword "desgraça".

News program	Twitter query	Twitter messages
JN	"Jornal Nacional só * desgraça"	**156**
JR	"Jornal da Record só * desgraça"	46

Tabela 3. Twitter search query including keywords "notícia", "ruim", "tragédia", "morte" and "violência".

News program	Twitter query	Twitter messages
JN	"jornal nacional só *" AND (notícia ruim OR tragédia OR morte OR violência)	**155**
JR	"jornal da record só * " AND (notícia ruim OR tragédia OR morte OR violência)	28

Preliminary results indicate that there is more criticism about JN than JR. In all, twitter users posted 311 messages criticizing the JN and 74 criticizing the JR. The number of messages criticizing JN is more than 400 percent greater than messages criticizing JR. These results can be explained by research which indicates that JN has an audience four times larger than JR [Ferreira, 2007].

7. Conclusion and Future Work

In this work, we used the LIWC dictionary to analyze words from news broadcasts. We analyzed Closed Captions from the two most watched Brazilian news broadcasts (i.e., JN and JR). Results indicate that JN CCs contain more negative words and positive emotions. In contrast, JR CCs contains more words related to negative emotions, anger, sadness and death.

We also analyzed Twitter posts criticizing JN and JR. Results reveal that there is much more criticism of JN than of JR. This can be explained by research which indicates that JN has a much larger audience. Thus, although news about crime and violence can increase the audience of a news channel, we confirmed that there are several criticisms of the negative content of Brazilian broadcast news.

Our future work will take advantage of the possibility of including other affective categories. Moreover, we intend to develop multimodal studies, correlating affective characteristics of words with facial analysis, in order to verify the emotion related to the facial expression in the process. We also expect to extend the analysis to other broadcast news channels.

Referências

Luanny Galvão Almeida. O descompasso entre a realidade midiática e a realidade processual e suas implicações para o julgamento criminal justo. *Revista Transgressões*, 5 (2):82–103, 2017.

Pedro Paulo Balage Filho, Thiago Pardo, and Sandra Aluísio. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In Sandra Maria Aluísio and Valéria Delisandra Feltrim, editors, *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 215–219, Fortaleza-CE, Brazil, 21–23 October 2013. Sociedade Brasileira de Computação.

Jacob Cohen. Statistical power analysis for the behavioral sciences 2nd edn, 1988.

Fernanda Vasques Ferreira. As representações dos indivíduos anônimos no telejornalismo brasileiro: um estudo comparativo entre o jornal nacional e o jornal da record. 2007.

Flavia Daniele Oliveira Gamonar. Planejamento e prototipagem de uma rede social de gastronomia convergente com programas de tv e mídias sociais. 2015.

Juliana Junqueira et al. Telejornalismo e violência urbana: cidadania nas notícias sobre criminalidade: realidade possível? 2013.

Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Mining social networks using heat diffusion processes for marketing candidates selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 233–242. ACM, 2008.

Li Ma, Zhong Tian Jia, Hai Yan Sun, and Chuan Yu. An improved model of the internet public opinion spreading on mass emergencies. In *Applied Mechanics and Materials*, volume 433, pages 1760–1764. Trans Tech Publ, 2013.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1): 547–577, 2003.

Jéssica Conceição Ribeiro. Sentidos sobre o jornalismo no twitter: uma análise do discurso dos integrantes sobre o jornalismo contemporâneo. 2015.

Ralph L Rosnow and Robert Rosenthal. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, 1(4):331, 1996.

Élida Mattos Vaz and Theresa Medeiros. Jornalismo e jornalistas na berlinda: Uma análise da abordagem negativa da imprensa e sua relação com a crise contemporânea da imprensa. In *XXXVII Congresso Brasileiro de Ciências da Comunicação*. Intercom, 2014.

Alfredo Pereira Vizeu. 65 anos de televisão: o conhecimento do telejornalismo e a função pedagógica/65 years of the television: the knowledge of the telejournalism and the pedagogical function. *Revista FAMECOS*, 23(3):1–17, 2016.