Análise de sentimento e envolvimento do público nas hashtags #VcNoJL1 e #VcNoJL2 no Twitter (X): um estudo de caso do telejornal TV Liberal no Pará, Brasil

Sentiment analysis and audience engagement in the hashtags #VcNoJL1 and #VcNoJL2 on Twitter (X): a case study of the TV Liberal newscast in Pará, Brazil

Alan Marcel Fernandes de Souza [a, *] iD

RESUMO: As transmissões televisivas em todo o mundo empregam *hashtags* específicas para facilitar o envolvimento do espectador através das redes sociais com os apresentadores dos programas. No contexto das transmissões de notícias, os comentários dos indivíduos misturam-se com as matérias exibidas nos telejornais, principalmente em relação às situações urbanas. Essas publicações nas redes sociais constituem uma base de dados significativa, uma vez que os usuários assumem papéis semelhantes aos dos cidadões. O objetivo deste artigo é esmiuçar as postagens associadas às *hashtags* #VcNoJL1 e #VcNoJL2, divulgadas na plataforma Twitter (X), no contexto do telejornal TV Liberal, veiculado no Estado do Pará (Brasil), utilizando a metodologia CRISP-DM e técnicas de ciência de dados. Empregando um exame estatístico nessas publicações, observou-se um nível discreto de envolvimento entre os telespectadores, que também são utilizadores da referida plataforma de redes sociais. Além disso, foi realizada uma avaliação de sentimento das postagens, revelando que 54% dos comentários são negativos, 31% são positivos e 15% são neutros. Os comentários negativos ressaltam notavelmente os desafios dentro da administração pública municipal, enquanto que os comentários positivos são em torno de ocasiões comemorativas como o Dia das Mães e o Dia dos Avós. Esse tipo de análise constitui uma importante ferramenta de monitoramento não só para o programa televiso, como também para a administração pública.

Palavras-chave: Telejornal; *Hashtags*; Interação ao Vivo; Ciência de Dados; Análise de Sentimento.

ABSTRACT: Television broadcasts around the world employ specific hashtags to facilitate viewer engagement via social media with program hosts. In the context of news broadcasts, individuals' comments mix with the articles shown on television news, mainly in relation to urban situations. These social media posts constitute a significant database, as users assume roles similar to those of citizens. The aim of this paper is to scrutinize the posts associated with the hashtags #VcNoJL1 and #VcNoJL2, published on the Twitter (X) platform, in the context of the TV news program TV Liberal, broadcast in the State of Pará (Brazil), using the CRISP-DM methodology and data science techniques. Using a statistical examination of these publications, a discreet level of engagement was observed among viewers, who are also users of the aforementioned social media platform. Furthermore, a sentiment assessment of the posts was carried out, revealing that 54% of the comments are negative, 31% are positive and 15% are neutral. The negative comments notably highlight the challenges within municipal public administration, while the positive comments are around commemorative occasions such as Mother's Day and Grandparents' Day. This type of analysis constitutes an important monitoring tool not only for the television program, but also for public administration.

Keywords: Television News; Hashtags; Live Interaction; Data Science; Sentiment Analysis.

[a] Department of Mathematics, Statistics and Informatics, Universidade Estadual do Pará, Castanhal, PA, Brasil.

* Correspondência para/*Correspondence to* Alan Marcel Fernandes de Souza . E-mail: alan.souza@uepa.br.

INTRODUCTION

Frequently, urban inhabitants assume the role of patrons in their interaction with services dispensed by governing entities (Adrees *et al* 2019; Alguliyev, Aliguliyev, Niftaliyeva 2019). This phenomenon becomes conspicuously pronounced when citizens utilize social media platforms as conduits to articulate their viewpoints concerning the administration of localities. Conversely, routine television broadcasts feature segments spotlighting predominantly instances of deficiencies in the oversight and implementation of initiatives undertaken by governmental functionaries, thereby exerting direct or indirect ramifications upon the populace's quality of life.

In contemporary times, the practice of news programs adopting specific hashtags and prompting viewers to engage in real-time interactions with program hosts has become ubiquitous (Rosa and Loureiro 2022). Such an approach not only functions as a means to stimulate and gauge viewer engagement, but also serves as a mechanism to approximate the viewership metrics of the pertinent news broadcast. Consequently, this facilitates the redirection of endeavors towards the immediate sustenance of the viewership base. Notably, the social media platform Twitter (recently referred to as X) traditionally assumes a pivotal role in fulfilling this communicative purpose.

The interface established between a television news program and a social media network provides a conduit for the systematic examination of messages concomitant with designated hashtags. This facilitates the utilization of data mining methodologies to unearth the extent of involvement exhibited by the viewership of the television program, along with discerning prevalent subjects of grievances or commendations, as encapsulated within sentiment analyses. Moreover, this symbiotic interaction serves to foster the phenomenon recognized as "citizen journalism," as expounded upon by Roberts (2019).

Guo and Liu (2022) investigated leveraged Twitter-derived data to comprehend the salient themes inherent within the online movements denoted by the hashtags #BlackLivesMatter and #StopAsianHate. Over a span of three months encompassing data collection, prominent topics encompassed instances of widespread violence, concerted communal endeavors, manifestations of community solidarity, and the articulation of censure against white racism. These significant aspects collectively constituted pivotal facets within the discourse surrounding the campaign against racial discrimination, which reverberated extensively across social media platforms.

Russel, Hernandes, Sanchez (2023) carried out an empirical investigation was undertaken to discern primary subjects to which Twitter users responded when prominent communication conglomerates engaged with diverse news content. The findings revealed that themes such as weather conditions and instances of criminal activity exhibited predictive potential for eliciting user-generated content sharing. However, in the context of sports-related content, a diminished propensity for garnering user approval through likes or retweets was evident.

Stewart (2019) conducted an examination of real-time Twitter posts pertaining to the sophomore season of the television show "The X Factor" in New Zealand. The study elucidated a discernible inclination among individuals to opt for live viewership of the program as a means to facilitate the dissemination of commentary. This practice fosters an interactive milieu amongst fellow viewers, effectively fostering the emergence of a community characterized by shared viewpoints pertaining to the program's content.

In Brazil, specifically within the confines of the State of Pará, resides a television station recognized as TV Liberal. This station maintains an affiliation with the commercial broadcasting network TV Globo, which commands a preeminent viewership across the nation (Brasil de Carvalho, 2019). TV Liberal airs a pair of local news broadcasts on a daily basis, excluding Sundays: the inaugural Jornal Liberal edition (JL1) airs at noon, followed by the subsequent Jornal Liberal second edition (JL2) at 7 pm. In both iterations of the program, hosts actively encourage public engagement through the utilization of social media platforms, employing the hashtags #VcNoJL1 and #VcNoJL2 respectively. Certain Twitter posts are meticulously selected and relayed live by the presenters as part of the broadcast.

The database formed by these posts remains pertinent, as the majority of individuals, in their commentary on news items, express their genuine and authentic opinions regarding the issues plaguing the city. This sincerity is inherent to the platform of social media, as comments are made through this medium rather than formalized forms or equivalent channels.

The objective of this paper is to quantitatively assess the degree of viewer engagement associated with Jornal Liberal, a program aired on TV Liberal, wherein audience interaction is facilitated via the employment of hashtags #VcNoJL1 and #VcNoJL2. This assessment encompasses both metric analysis and sentiment evaluation embedded within the posted content. To this end, the subsequent research questions (RQ) have been formulated: RQ#1: When performing a statistical evaluation of post metrics, what is the engagement level of posts? RQ#2: What is the prevailing sentiment encapsulated in the posts and how is it identified and characterized?

Section 2 of this article delineates the employed methodology, elucidating the approach undertaken. Subsequently, Section 3 presents the outcomes and ensuing discourse, providing an in-depth examination and interpretation of the findings. Lastly, Section 4 encapsulates the article's conclusion, summarizing the key takeaways derived from the study.

METHODOLOGY

To achieve the proposed objective, the Cross Industry Standard Process - Data Mining (CRISP-DM) methodology was used because it is well-known, well-documented, and divided into flexible phases (Schröer, Kruse, Gómez 2021; Garcia-Arteaga et al 2024; Guzmán Ordóñez et al 2023). The first phase of the methodology was understanding

the business, where the objective was to identify that the news program continuously establishes and promotes hashtags in each of its editions: #VcNoJL1 and #VcNoJL2.

The second phase consisted of understanding the data, verifying the necessity to automatically extract Twitter posts due to the large volume of data. This acquisition process was systematically executed via automated web scraping techniques, utilizing the Selenium tool (version 4.8.3). This tool proved efficacious in data extraction, obviating the necessity of utilizing a paid Application Programming Interface (API). Moreover, according Dongo *et al* (2020), this approach is faster than using Twitter API to access the dataset. The data collection interval spanned from March 18, 2023, to July 31, 2023, spanning a duration of 135 days, yielding a total of 2,367 extracted posts. Web scraping is a technique for automatically extracting data from websites. By using or building computer programs, known as scrapers or bots, it is possible to access web pages, extract relevant content, and store it in a structured format, such as tables, databases, or CSV (comma-separated values) files, without human intervention (Khder 2021; Singrodia, Mitra, Paul 2019).

The third phase took into account data preparation. When the posts were extracted directly from Twitter platform, the data was stored in a CSV file. The columns (variables) generated in this file were: hashtag, date of posting, post content, post sentiment class, sentiment score, post responses count, reposts count, likes count, and views count. This phase also encompassed data pre-processing, entailing the elimination of duplicate entries, appropriate formatting of distinct data types (including integers, real numbers, timestamps, and textual content), and the removal of null or incongruent data instances. This process facilitated the creation of a meticulously organized CSV file, poised for utilization in the ensuing phase of the methodology.

It is observed that the sentiment classification of each post was done in one of the following classes: negative, neutral, and positive. To accomplish this, the IBM Watson Natural Language Processing SDK (Software Development Kit) version 1, which was launched on April 7, 2022, was employed. According to Solanki (2022), the artificial intelligence technique used by IBM Watson to classify the sentiment of a text is BERT (Bidirectional Encoder Representations for Transformers). It is a widely used technique for this purpose and generates accurate results (Alaparthi, Mishra 2021; Kaur 2022). This natural language processing (NLP) and sentiment analysis of the posts is part of the fourth phase of the CRISP-DM used in this study: modeling.

After that, the fifth phase consisted of evaluating the results generated by IBM Watson: in conjunction with the sentiment classification, the IBM Watson system computed a sentiment score within the range of [-1, 1]. Instances wherein values fell below zero were categorized as negative sentiments, whereas those exceeding zero were deemed positive, and a score of zero corresponded to a classification of neutral sentiment.

The sixth phase, which is the implementation of the solution, was carried out through the computation of descriptive statistical measures for the dataset, in tandem with the

application of data and text mining techniques. These combined analytical approaches facilitated the formulation of responses to the pre-established research inquiries.

The ensuing sections expound upon the elucidation of these findings. The figure 1 summarizes the data science project cycle applied in this study and which was based on the CRISP-DM methodology.

Figure 1. Data science project cycle used in this study.
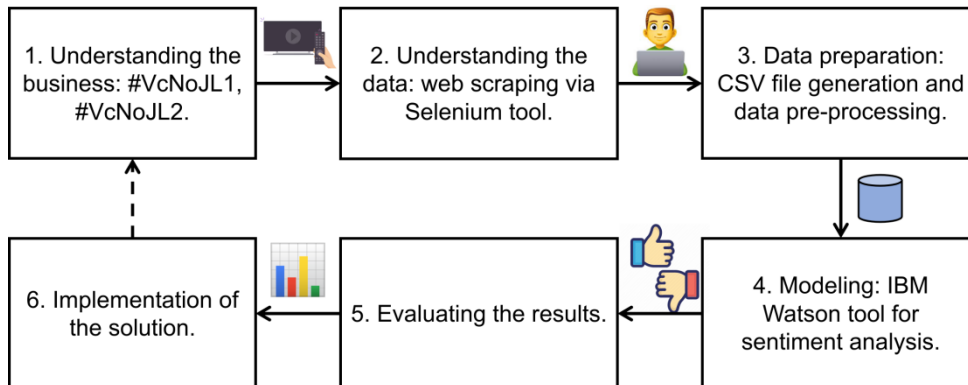

RESULTS AND DISCUSSION

As mentionated in Section 2, a total of 2,367 posts were amassed during the 135-day data collection period, culminating in an average daily count of 17.53 posts. Table 1 provides an overview of the descriptive statistical metrics pertaining to the quantitative attributes (quantities) of the posts. Notably, the mean values for post responses, reposts, and likes all fall below unity. Additionally, the percentiles uniformly manifest zero values, underscoring a conspicuous paucity of interaction among Twitter users themselves. Conversely, the view count assumes comparatively larger values; however, their significance is diminished in relation to the other metrics, as mere message viewing demands minimal "engagement" on the part of the user, confined to perusal within their timeline. It is also clear, through this statistical analysis, that there are no digital influencers involved in the posts and public engagement in question, since the standard deviation is small.

Table 1. Descriptive statistics of post metrics.

Parameter	Replies	Reposts	Likes	Views
Average	0.054	0.033	0.362	42.355
Standard deviation	0.279	0.230	1.472	107.945
Minimum	0	0	0	1
Maximum	3	5	44	4091
25% percentile	0	0	0	17
50% percentile	0	0	0	25
75% percentile	0	0	0	39

The graph depicted in Figure 2 was formulated to discern the distribution of posts across different days of the week. Notably, Tuesday, Wednesday, and Thursday emerge as the days characterized by heightened user interaction, marked by a surge in post frequency. It is noteworthy to observe that despite the absence of television news broadcasts on Sundays, a noteworthy accumulation of over 100 posts is discernible, which are anticipated to be referenced on the subsequent Monday.

Figure 2. Posts amount by week day.
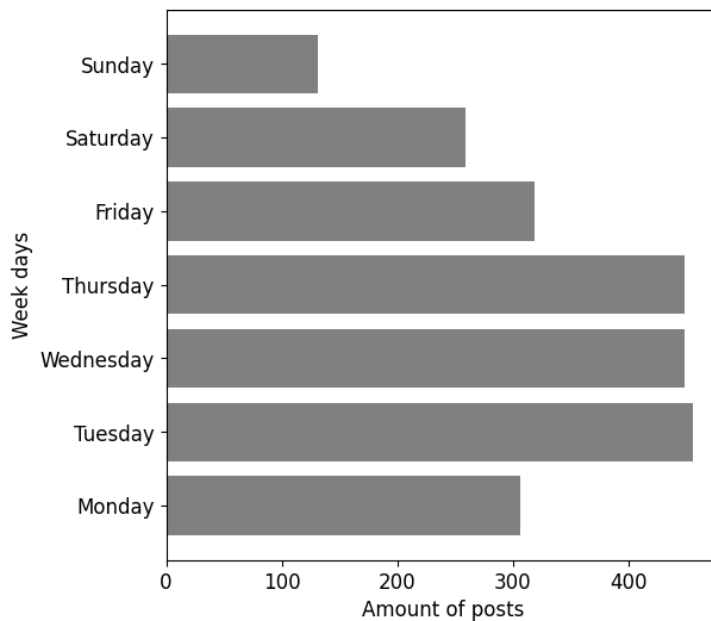
An additional analysis conducted encompassed the categorization of posts based on their associated sentiments. The majority of posts (54%) were assigned a negative sentiment, while 31% exhibited a positive sentiment, and the remaining 15% were characterized as neutral. Illustrated in Figure 3 is a discernible imbalance, with a heightened occurrence of negative comments in contrast to positive ones. It is worth highlighting two noteworthy points in the graph: around the 1000th post mark and approximately the 2600th post, concentrations of positive comments become evident. These specific junctures correspond to posts related to Mother's Day and Grandparents' Day, respectively. On these commemorative occasions, a notable number of individuals frequently share photos accompanied by positive messages for live broadcast during the news program, thereby paying homage to their families.

Figure 3. Dispersion of posts' sentiments.

With respect to the negative posts, an examination of the word cloud presented in Figure 4 elucidates recurring themes of discontent. A prevalent issue pertains to the deficiency of particular services within the city. Furthermore, a notable number of individuals voice grievances concerning the suboptimal municipal management undertaken by the local administration. Complaints regarding unsatisfactory waste management practices, exemplified by concerns related to trash, and the prevailing chaotic traffic conditions also surface prominently. To a lesser degree, critiques extend to the water distribution service, the state of educational institutions, instances of violence, and inadequacies within the public transportation system. Notably, the posts were in Brazilian Portuguese. To generate the word cloud for this article, the words were translated into English literally, with no ambiguous words being identified.

Figure 4. Word cloud of posts.

Liinc em Revista, Rio de Janeiro, v. 20, n. 01, e7033, jun 2024.

https://doi.org/10.18617/liinc.v20i1.7033

7/11

Frame 1 displays examples of negative, positive and neutral comments randomly chosen from the database.

Frame 1. Posts examples by sentiment classes.

Sentiment	Post (translated)
Negative	a) "Regrettably, these rulers what they are doing with Belém are filthy fairs, dirty streets a shame unfortunately how to trust these politicians the example is there visible to all #VCnoJL1" b) "The outburst of the citizen of Parauapebas was very impactful, it is not just about precariousness in education - it is as if the neighborhood was forgotten. In fact, we feel that way in the 4 corners of Pará. Desperate reality. #VCnoJL2" c) "#VCnoJL1 Good afternoon Priscila Belém is abandoned Mayor, one worse than the other when will we have a mayor who loves our city and takes care of it...?"
Positive	d) "Happy mother's day queen, you are amazing, despite everything you've been through, you're still a warrior!! I love you (and a happy mother's day to all who accompany JL1) #VCnoJL1" e) "#vcnojl2 can't wait for the Círio 2023, this one is going to be good!" f) "#VCnoJL2 I would like to congratulate the governor on the beautiful facade of Ophir Loyola, but patients are still without medication and lack of beds for surgery."
Neutral	g) "Belém should copy Curitiba's model in terms of public transport and bus stops. #VCnoJL1" h) "#VCnoJL1 It is common to observe a large number of cyclists risking their lives traveling freely on the BRT and there is no supervision or guidance service on the part of traffic agents." i) "@JL2edicao The partnership between the prefecture and the private sector could solve the problem of bus stops in the capital. #VCnoJL2"

This dissatisfaction with the (in)actions of public authorities has also been addressed by Cabral Filho and Oliveira (2022), who identified that there are people who seek visibility for local causes, on the internet, in contrast to the attentional dynamics established in the public sphere.

Besides, sentiment analysis has already been used by Malini, Ciarelli, Medeiros (2017) to measure emotions on Twitter regarding the impeachment of Dilma Roussef, former president of Brazil, identifying feelings consistent with the reality of that moment. Afonso (2017) also used sentiment analysis to analyze the feelings extracted from referencing YouTube texts.

CONCLUSION

The customary practice of sharing live television show links on Twitter engenders a hybrid communication conduit. This channel fosters multifaceted interaction, encompassing viewers who contribute posts, program hosts, and the wider audience. Consequently, this dynamic configuration facilitates the quantification of engagement levels and audience reach for the programs, concurrently affording insights into the prevailing sentiment encapsulated within the posted commentaries.

Based on the comprehensive overview delineated in Section III, the research questions posed can be addressed as follows:

For RQ#1: When performing a statistical evaluation of post metrics, what is the engagement level of posts?

In response to RQ#1: The evaluation of post metrics signifies an exceedingly low level of engagement, as deduced from the statistical analysis encompassing metrics such as post responses, reposts, likes, and comment views. Perhaps relying on digital influencers from Twitter could improve engagement.

For RQ#2: What is the prevailing sentiment encapsulated in the posts and how is it identified and characterized?

In response to RQ#2: The dominant sentiment prevalent within the posts, as classified by IBM Watson, is predominantly negative, accounting for 54% of the total. Utilizing the word cloud visualization, a prominent recurring theme involves commentary pertaining to the broadcast reporters names (Priscila and João), consistently highlighting the challenges afflicting the city of Belém capital of Pará located Brazil (frequent terms by the word cloud presented by Figure 3: prefecture, trash, missing, sad, city). These observations are consistently corroborated by the public via their posted expressions.

Studies of this nature bear significant importance due to the following reasons:

(A) They represent a contemporary and innovative approach to gauging the engagement and reception exhibited by consumers of television news.

(B) In the context of television news, the scrutiny of sentiment within posts serves as an effective barometer for reflecting the city's challenges. This reservoir of information proves invaluable for administrators, offering insights into citizen grievances, thereby facilitating targeted enhancements.

(C) By incorporating live post monitoring and analysis software throughout the program's duration, the potential for live adjustments to the program script arises. This dynamic adaptation strategy holds promise in sustaining and retaining the audience's interest. It is important to public administration too, once data may be collected freely.

DATA AVAILABILITY

The entire dataset and the software developed and used that support the results of this study were made available on Github and may be accessed at https://github.com/amarcel/scrap-twitter-vcnojl.

REFERENCES

ADREES, M. S., et al. A Framework of Promoting Government Services using Social Media: Sudan E-Government Case Study. *Journal of Physics: Conference Series* [online]. 2019, **1167**, 012062 [viewed 6 May 2024]. ISSN 1742-6596. Available from: doi:10.1088/1742-6596/1167/1/012062

AFONSO, Alexandre Ribeiro. A referenciação em textos do YouTube: um estudo com vistas à análise de sentimentos. *Liinc em Revista* [online]. 2017, **13**(2) [viewed 6 May 2024]. ISSN 1808-3536. Available from: doi:10.18617/liinc.v13i2.3933

ALAPARTHI, Shivaji, and Manit MISHRA. BERT: a sentiment analysis odyssey. *Journal of Marketing Analytics* [online]. 2021, **9**(2), 118–126 [viewed 9 July 2024]. ISSN 2050-3326. Available from: doi:10.1057/s41270-021-00109-8

ALGULIYEV, Rasim M., Ramiz M. ALIGULIYEV, and Gunay Y. NIFTALIYEVA. Extracting social networks from e-government by sentiment analysis of users' comments. *Electronic Government, an International Journal* [online]. 2019, **15**(1), 91 [viewed 6 May 2024]. ISSN 1740-7508. Available from: doi:10.1504/eg.2019.096576

BRASIL DE CARVALHO, Vanessa. Ciência em publicidades: uma análise das emissoras televisivas de maior audiência no Brasil. *Chasqui. Revista Latinoamericana de Comunicación* [online]. 2019, **1**(140), 297–314 [viewed 6 May 2024]. ISSN 1390-924X. Available from: doi:10.16921/chasqui.v0i140.3576

CABRAL FILHO, Adilson Vaz, and Cinthya Pires OLIVEIRA. Democracia em vertigem: reflexões sobre os fluxos de atenção e ações coletivas para mudança social na Internet. *Liinc em Revista* [online]. 2022, **18**(2), e6052 [viewed 6 May 2024]. ISSN 1808-3536. Available from: doi:10.18617/liinc.v18i2.6052

DONGO, Irvin, et al. Web Scraping versus Twitter API. In: *iiWAS '20: The 22nd International Conference on Information Integration and Web-based Applications & Services* [online]. New York, NY, USA: ACM, 2020 [viewed 9 July 2024]. Available from: doi:10.1145/3428757.3429104

GARCIA-ARTEAGA, Joel, et al. An effective approach for identifying keywords As high-quality filters to get emergency-implicated Twitter Spanish data. *Computer Speech & Language* [online]. 2023, 101579 [viewed 9 July 2024]. ISSN 0885-2308. Available from: doi:10.1016/j.csl.2023.101579

GUZMÁN ORDÓÑEZ, Anabel, et al. Analytical model to measure the effectiveness of content marketing on Twitter: the case of governorates in Colombia. *Journal of Marketing Analytics* [online]. 2023 [viewed 9 July 2024]. ISSN 2050-3326. Available from: doi:10.1057/s41270-023-00243-5

KAUR, Parneet. Sentiment analysis using web scraping for live news data with machine learning algorithms. *Materials Today: Proceedings* [online]. 2022 [viewed 9 July 2024]. ISSN 2214-7853. Available from: doi:10.1016/j.matpr.2022.05.409

KHDER, Moaiad. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications* [online]. 2021, **13**(3), 145–168 [viewed 9 July 2024]. ISSN 2074-8523. Available from: doi:10.15849/ijasca.211128.11

SOLANKI, Shivam. IBM Developer. *IBM Developer* [online]. 22 October 2022 [viewed 9 July 2024]. Available from: https://developer.ibm.com/tutorials/use-the-watson-core-nlp-library-to-perform-sentiment-analysis/

GUO, Jing, and Shujun LIU. From #BlackLivesMatter to #StopAsianHate: Examining Network Agenda-Setting Effects of Hashtag Activism on Twitter. *Social Media + Society* [online]. 2022, **8**(4), 205630512211461 [viewed 6 May 2024]. ISSN 2056-3051. Available from: doi:10.1177/20563051221146182

MALINI, Fabio, Patrick CIARELLI, and Jean MEDEIROS. O sentimento político em redes sociais: big data, algoritmos e as emoções nos tweets sobre o impeachment de Dilma Rousseff. *Liinc em Revista* [online]. 2017, **13**(2) [viewed 6 May 2024]. ISSN 1808-3536. Available from: doi:10.18617/liinc.v13i2.4089

ROBERTS, Jessica. Citizen journalism. *The International Encyclopedia of Media Literacy.* 2019, 1-10, [viewed 6 May 2024]. Available from: https://www.researchgate.net/profile/Jessica-Roberts-3/publication/333111582_Citizen_Journalism/links/5cf137d1299bf1fb184d7b3d/Citizen-Journalism.pdf.

ROSA, Antonio Celestino, and Marcos Dantas LOUREIRO. O uso de hashtags e a gestão algorítmica de dados no Instagram. *Liinc em Revista* [online]. 2022, **18**(2), e6054 [viewed 6 May 2024]. ISSN 1808-3536. Available from: doi:10.18617/liinc.v18i2.6054

RUSSELL, Frank M., Miguel HERNANDEZ, and Korryn SANCHEZ. #BREAKING local news: Twitter use in a large media market. *Newspaper Research Journal* [online]. 2023, 073953292211433 [viewed 6 May 2024]. ISSN 2376-4791. Available from: doi:10.1177/07395329221143376

SCHRÖER, Christoph, Felix KRUSE, and Jorge Marx GÓMEZ. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science* [online]. 2021, **181**, 526–534 [viewed 9 July 2024]. ISSN 1877-0509. Available from: doi:10.1016/j.procs.2021.01.199

SINGRODIA, Vidhi, Anirban MITRA, and Subrata PAUL. A Review on Web Scrapping and its Applications. In: *2019 International Conference on Computer Communication and Informatics (ICCCI)* [online]. IEEE, 2019 [viewed 9 July 2024]. ISBN 9781538682609. Available from: doi:10.1109/iccci.2019.8821809

STEWART, Mark. Live tweeting, reality TV and the nation. *International Journal of Cultural Studies* [online]. 2019, **23**(3), 352–367 [viewed 6 May 2024]. ISSN 1460-356X. Available from: doi:10.1177/1367877919887757