Addressing Database-Related Issues in Digital Social Network Data Analysis

Alexandre Heine[1], Mariana Porto Barreto[1], Rodrigo Motta[1], Edward Hermann Haeusler[1], Sérgio Lifschitz[1]

[1]Departamento de Informática – (PUC-Rio) - Rio de Janeiro - RJ

{aheine,mbarreto,rferreira,hermann,sergio}@inf.puc-rio.br

Abstract. *The analysis of vast Digital Social Network datasets often neglects database systems, compromising effectiveness and efficiency, particularly in qualitative studies. We present some preliminary practical experiments, including quantitative mention counting and qualitative topic identification, showcasing (i) that managing large-scale social data is relevant and (ii) how leveraging DBMS could yield more robust and insightful social network data analysis.*

1. Introduction

Communication through a Digital Social Network (DSN) is part of people's daily lives. Because of this, observing and analyzing the data and information circulating through DSNs has become relevant, as the online behavior of society can reflect the thoughts, opinions, and stances of community members in their offline routines.

DSNs are frequently mentioned as canonical examples of Big Data systems. In the broader field of computer science, in the presence of large quantities of data, the use of computational systems employing Database Management Systems (DBMS) is strongly recommended. However, the literature review on DSNs curiously shows that database systems are rarely used. In this work, we specifically address the effective use of database systems for managing data originating from DSNs. We seek to motivate using database systems for investigations in the area. In some relevant and often frequent cases, we claim that not considering significant data volumes, with or without DBMS, can lead to incomplete, incorrect, or even inconclusive analyses.

The manuscript is organized as follows: after this brief introduction, Section 2 motivates our research while Section 3 presents a literature review, emphasizing the amount of data and the systems considered. In Section 4 we propose and conduct two selected experiments on actual DSN data regarding national elections in Brazil. We conclude this article in Section 5 with an evaluation of the results and future work.

2. Motivation

The research reported in this work concerns managing and handling large volumes of data obtained from applications that support Digital Social Networks (DSNs), such as X (formerly Twitter), Facebook, TikTok, Reddit, etc. Any research work in this area somehow involves defining a *data-driven computational pipeline*, eventually complemented by a *human-in-the-loop* step. This pipeline begins with data collection (e.g., using available APIs), cleaning and filtering this data to remove unnecessary, incorrect, or incomplete entries, and enriching the data with external sources followed by data modeling to ensure efficient management and access. With the data properly prepared and organized, various analyses – from simple statistics to complex ones involving DSN-specific knowledge – can be performed, visualized appropriately, or exported to specialized user tools.

DSNs are always mentioned when referring to Big Data systems because they involve large volumes of data generated by streaming and with great variety, to name just a few of the Vs that typically characterize these systems. Due to the volume of data, DSNs are expected to be associated with well-known database systems, in various models (e.g., relational, document, or graph). However, a review of the specialized literature surprisingly shows that research works rarely cite the use of any DBMS to handle the large quantities of collected data. Does this imply that, in practice, *DBMSs are unnecessary because we do not need such large datasets after all?*

And we can go further: even when datasets are voluminous, only these data samples are considered during analysis. This means such samples must be *adequately significant* to avoid biasing analyses or even generating undue and incorrect conclusions. Given this context, we believe it is essential to be clear about which types of problems we can address with significant data samples and in which other situations, more complete and larger datasets and database systems are needed.

3. Related Works

For this particular research, we decided to investigate the publications from two prominent Brazilian events dealing with DSNs in detail: the Brazilian Workshop on Social Network Analysis and Mining (BRASNAM) and the Brazilian Symposium on Databases (SBBD). For BRASNAM, across all its occurrences since 2012 and SBBD from 2016 on, for all available publications at the SBC SOL portal[1].

We have checked 83 articles published at BRASNAM and 17 at SBBD. We must mention that most articles (71%) considered X (Twitter) as the primary dataset, when compared to Facebook, Instagram, Reddit, and TikTok, to name a few. This prominence is mainly due to it being one of the oldest social networks still in activity and, crucially, for historically offering more accessible mechanisms for data collection. Among other collected information, we were interested in the data volumes (e.g., quantity of posts) obtained, the method of data capture (e.g. API, crawling or scraping), and whether the conducted research considered any system supported by some database system or if they directly used spreadsheet systems for file handling, such as CSV or Parquet.

Considering the BRASNAM and SBBD papers, 19 out of 100 published articles do not comment on the quantity of data involved. For manuscripts that mention something about the data, 40% considered a maximum of 100,000 collected posts, and all the others deal with at most 1,000.000 posts. Articles that cite volumes exceeding 10 million captured data from various social networks are rare. And even the most voluminous ones, as was the case with [Gonçalves et al. 2012], mostly mention having performed analyses on samples of the total set.

Also, a few articles, including these 13 years of the BraSNAM event, mention whether they are using any system to deal with large volumes of data (e.g., [Carmo et al. 2023]), or even specify which DBMS-like system, relational or not,

[1] https://sol.sbc.org.br/index.php/indice

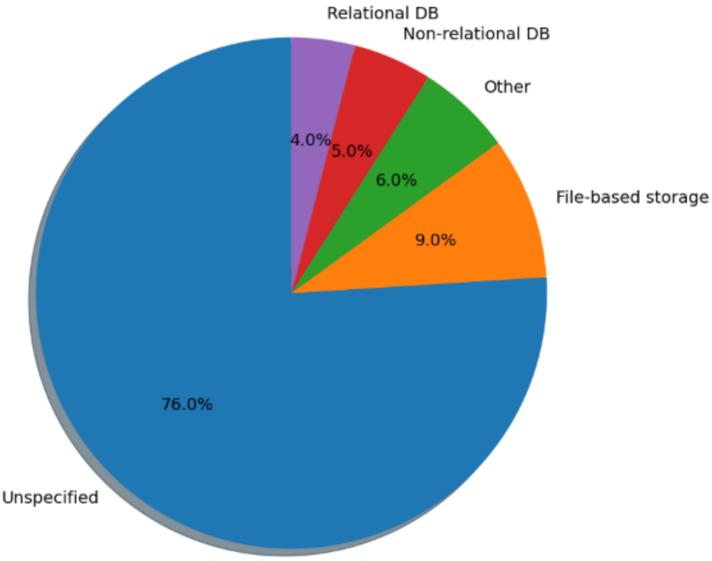


Figure 1. Overview of Related Works

they are using. Not more than three articles mention the MongoDB system (a JSON-oriented NoSQL document store) (e.g., [Maia et al. 2021, Moreira et al. 2015, Costa et al. 2022, Rodrigues et al. 2024]. Other works reference the key-value NoSQL store Redis [Alves et al. 2022], or the relational DBMS PostgreSQL [Assis et al. 2015] and MySQL[Cunha and Gomes 2020]. Several publications explicitly mention data files in CSV format, but it is unclear if the use was for the persistence of collected data or solely for making the dataset available. Some authors generally mention the use of database systems or file storage systems, but even then, this number does not exceed 25.

Due to space limitations, we could not refer here to a broader list of articles, including recent published technical papers (e.g. [Alizadeh et al. 2024],[Singh et al. 2024] or [Singh et al. 2025]). However, dealing with large volumes is neither discussed nor discussed rarely in this literature. Due to space limitations, we could not consider more venues like international conferences and journals. However, our preliminary literature review indicates we may obtain similar results for a wider collection.

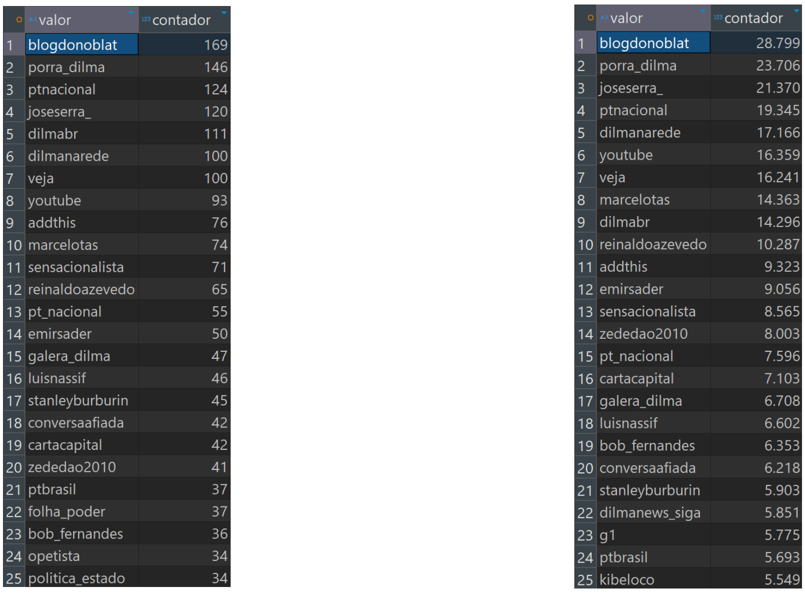
4. Experiments and Practical Evaluation

Our claims in this paper are illustrated with two different case studies considering posts about Brazilian elections: first, a quantitative example about DSN handles (accounts), and second, a qualitative example about the mentions of DSN handles (accounts). Then, a rather qualitative study involving topic modeling. For reproducibility issues, all our collected data and running examples are available at a GitHub repository: https://github.com/Maxteralex/representative-dataset-problem.

Case 1 - Mention Accounting: In this particular study, we consider a dataset of tweets from X concerning the second round of the 2010 Brazilian presidential elections with 3,105,247 posts. To generate a sample from this dataset, the formula for calculating sample size by proportions was used [Israel 1992]. This formula, as shown in Equation 1, calculates the necessary size for a sample (n_0) in an infinite population, based on the margin of error, the Z-score (obtained from a table based on the desired confidence level), and the degrees of variability ”p” and ”q,” which are set to 0.5 to maximize the sample size.

$$n_0 = \frac{Z^2pq}{e^2} \qquad (1) \qquad\qquad n = \frac{n_0}{1 + \dfrac{n_0 - 1}{N}} \qquad (2)$$

For an infinite population, a Z-score of 2.575 was obtained for a 99% confidence level, a margin of error of ±1%, and variability degrees (p and q) of 0.5. Consequently, the calculated sample size (n_0) is 16,577. To obtain a sample for a specific number of posts, the proportion correction formula (Equation 2) is used. To achieve a sample that meets the desired margin of error and confidence level, we need a sample of at least 16,489 posts for the 2010 dataset and 16,577 posts for the 2022 dataset. For these particular experiments, 19,000 random tweets were used as samples to meet the minimum requirements for both cases.



	valor	contador
1	blogdonoblat	169
2	porra_dilma	146
3	ptnacional	124
4	joseserra_	120
5	dilmabr	111
6	dilmanarede	100
7	veja	100
8	youtube	93
9	addthis	76
10	marcelotas	74
11	sensacionalista	71
12	reinaldoazevedo	65
13	pt_nacional	55
14	emirsader	50
15	galera_dilma	47
16	luisnassif	46
17	stanleyburburin	45
18	conversaafiada	42
19	cartacapital	42
20	zededao2010	41
21	ptbrasil	37
22	folha_poder	37
23	bob_fernandes	36
24	opetista	34
25	politica_estado	34

	valor	contador
1	blogdonoblat	28.799
2	porra_dilma	23.706
3	joseserra_	21.370
4	ptnacional	19.345
5	dilmanarede	17.166
6	youtube	16.359
7	veja	16.241
8	marcelotas	14.363
9	dilmabr	14.296
10	reinaldoazevedo	10.287
11	addthis	9.323
12	emirsader	9.056
13	sensacionalista	8.565
14	zededao2010	8.003
15	pt_nacional	7.596
16	cartacapital	7.103
17	galera_dilma	6.708
18	luisnassif	6.602
19	bob_fernandes	6.353
20	conversaafiada	6.218
21	stanleyburburin	5.903
22	dilmanews_siga	5.851
23	g1	5.775
24	ptbrasil	5.693
25	kibeloco	5.549

Figure 2. Mentions: Sample vs Complete Set

We compare a statistical analysis, specifically the number of mentions made to a particular user, either in a sample or the complete dataset. For this, the 2010 presidential election second-round dataset and the 19,000 random tweet sample were used to verify if the query returning the 25 most mentioned users would show little variation. Figure 2 left and right demonstrates that the top 25 remained largely consistent. Only users @dilmanews_siga, @g1 and @kibeloco were absent from the sample, replaced by @folha_poder, @opetista and @politica_estado. Regarding the order, most users shifted up or down by only one or two positions between the two scenarios. We conclude that the sampling was successful for this statistical analysis; we achieved the same results as the complete dataset using a significantly smaller amount of data.

Case 2 - Clusters and Topic Modeling: In this second case, we consider a dataset of tweets from X, considering the first day of the second round in the 2022 Brazilian presidential elections, with 1,291,891 posts. The complete second round would sum to over 60 million posts. For our example, we did not need a larger dataset.

Again, we consider a sample of 19,000 random tweets and the complete dataset

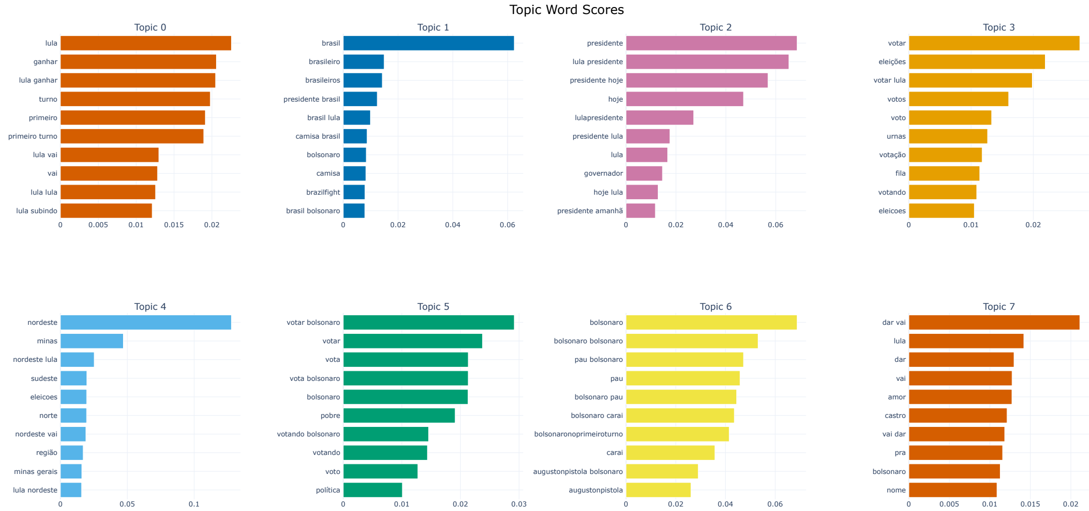


Figure 3. Topic clustering and detailed results for the dataset sample.

of the first voting day in the 2022 Brazilian presidential elections. This experiment aimed to determine if training a topic modeling model on the sample would yield a similar topic distribution for the complete dataset. BERTopic [Grootendorst 2022], a topic modeling framework, was employed to achieve this. BERTopic uses text embeddings to create clusters that represent groups of words, which in turn represent each topic. Using BERTopic, a topic identification model was trained on the sample, resulting in the topic distribution shown on Figure 3. This trained model was then used to classify the complete dataset, which in turn generated Figure 4. This figure clearly shows that the distribution of these topics in the complete dataset differs significantly from what was observed using only the sample.

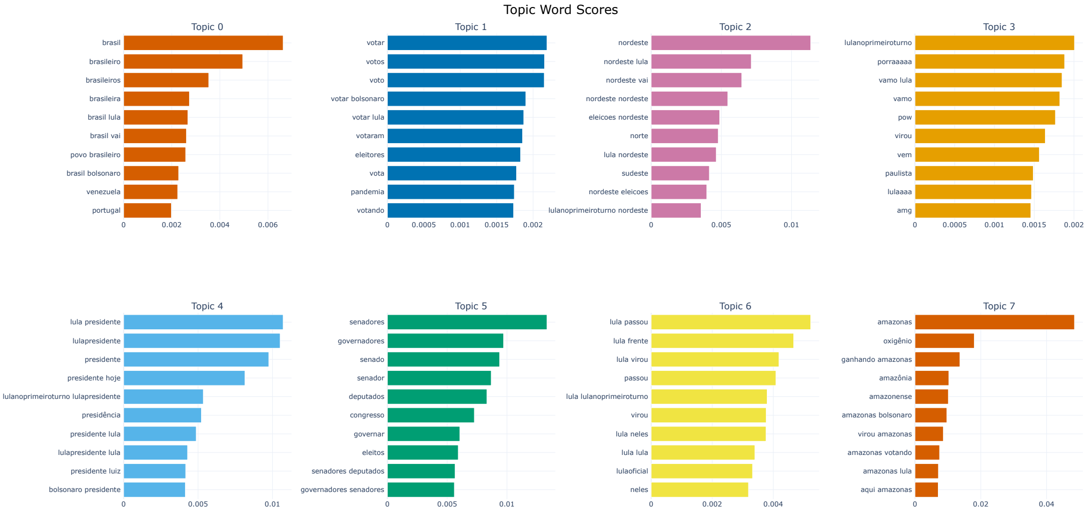


Figure 4. Topic clustering and detailed results for the complete dataset.

These results on topic modeling may also be further checked on Figure 5, where we may see on the left (sample dataset) all the colors from topics completely mixed up within two *clusters* that cannot help identify those major topics, while on the right the colors are correctly grouped, enabling the identification of topic communities.

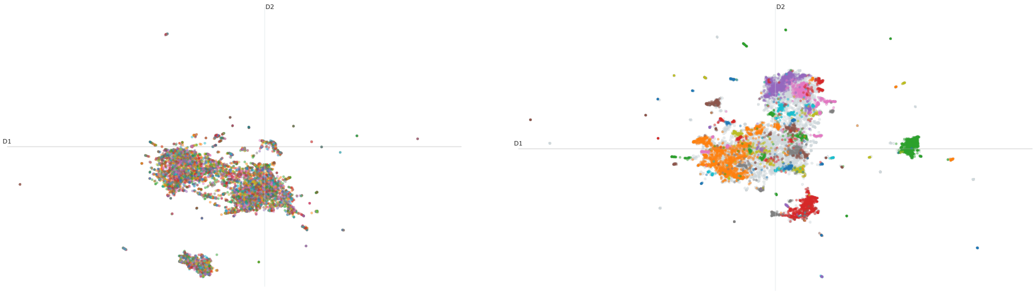


Figure 5. Topic Clustering: Sample vs Complete Set

We can conclude that for more qualitative analyses, such as this case of community mapping (or clusters) aiming to identify topics discussed during an electoral period, the use of samples can lead to biased results and provoke incorrect conclusions about the behavior of online and offline communities.

5. Conclusions

We have explored a highly relevant research question within the context of DSNs: how to handle large volumes of data collected from platforms so that some analyses are neither biased nor incomplete. Systems should manage effectively and efficiently the large volumes of data obtained from DSN platforms. According to our preliminary investigations, this trend is also repeated in articles published in other events or journals in the field.

In practice, it's observable that authors and researchers consider samples of existing data to perform their analyses due to the difficulty of handling voluminous datasets. However, determining significant samples for specific contexts, particularly for Big Data systems, is a very complex problem [Liu and Zhang 2020]. In this article, we chose two widespread types of analyses on DSNs and showed that even though there are cases (counting mentions) where traditional statistical samples work satisfactorily, there are other situations (communities and topic modeling) where it becomes necessary to consider larger and more complete datasets than simple samples.

As future work, some already underway, we intend to research how to adapt datasets obtained from DSN platforms for database systems (SQL or NoSQL) so that at least the problem of dealing with large volumes is mitigated. This research involves defining appropriate data models, independent of currently available technologies. With a proper conceptual definition, like the one we presented in this article, it's also possible to implement ad hoc analysis functionalities in the DSN domain that meet usual requirements and metrics, such as topic modeling, but also searches involving the evaluation of posts that went viral or generated greater engagement.

References

Alizadeh, M., Zare, D., Samei, Z., Alizadeh, M., Kubli, M., Aliahmadi, M., Ebrahimi, S., and Gilardi, F. (2024). Comparing methods for creating a national random sample of twitter users. *Social Networks Analysis and Mining*, 14(1):160.

Alves, O., Falcão, T., Valença, G., and Andrade, E. (2022). Brimo: uma ferramenta para análise de sentimentos. In *Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*, pages 97–108. SBC.

Assis, L., Horita, F., Herfort, B., Steiger, E., and de Albuquerque, J. (2015). Uma abordagem geográfica para a priorização de mensagens de mídias sociais para o gerenciamento de risco de inundação com base em dados de sensores. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*. SBC.

Carmo, I., Rêgo, A. L. C., Barreto, M., Schuler, M., Heine, A., Villas, M., and Lifschitz, S. (2023). Social networks managing with network analysis and topic modeling (in portuguese). In *SBBD Extended Proceedings*, pages 64–70.

Costa, G., Couto, D., Junior, A. J., and Lobato, F. (2022). Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*, pages 169–180. SBC.

Cunha, K. and Gomes, A. (2020). Mensuração do capital social acumulado a partir de interações sociais em páginas institucionais no facebook. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*, pages 85–96. SBC.

Gonçalves, P., Dores, W., and Benevenuto, F. (2012). Panas-t: Uma escala psicométrica para medição de sentimentos no twitter. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*, pages 153–164. SBC.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure, ArXiv. https://arxiv.org/abs/2203.05794.

Israel, G. D. (1992). Determining sample size. *Univ. of Florida Coop. Extension Service*.

Liu, Z. and Zhang, A. (2020). Sampling for big data profiling: A survey. *IEEE Access*, 8:72713–72726.

Maia, M., Oliveira, E., and Gallegos, L. (2021). Covid-19 e tweets no brasil: coleta, tratamento e análise de textos com evidências de estados afetivos alterados em momentos impactantes. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*, pages 79–90.

Moreira, T., Valerio, J., and Oliveira, J. (2015). Análise da confiabilidade da informação propagada em mídias sociais. In *Braz. Ws. Soc. Net. Analysis and Mining (BRASNAM)*.

Rodrigues, E., Pires, C. E., and Filho, D. N. (2024). Schema incremental evolution for document-oriented databases. In *SBBD*, pages 260–273.

Singh, S. S., Muhuri, S., Mishra, S., Srivastava, D., Shakya, H. K., and Kumar, N. (2024). Social network analysis: A survey on process, tools, and application. *ACM Comput. Surv.*, 56(8):192:1–192:39.

Singh, S. S., Singh, S., Singh, K., Srivastava, V., and Shakya, H. K. (2025). Big data meets social networks: A survey of analytical strategies and research challenges. *IEEE Access*, 13:98668–98698.