Universidade Federal da Bahia
Instituto de Matemática e Estatística

Programa de Pós-Graduação em Ciência da Computação

TEMPORAL NOVELTY QUANTIFICATION: A NEW APPROACH TO QUANTIFY NOVELTY IN SOCIAL NETWORKS

Victor Maciel Guimarães dos Santos

DISSERTAÇÃO DE MESTRADO

Salvador
30 de julho de 2020

VICTOR MACIEL GUIMARÃES DOS SANTOS

TEMPORAL NOVELTY QUANTIFICATION: A NEW APPROACH TO QUANTIFY NOVELTY IN SOCIAL NETWORKS

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ricardo Araújo Rios

Salvador
30 de julho de 2020

Sistema de Bibliotecas - UFBA

dos Santos, Victor Maciel Guimarães.

Temporal Novelty Quantification: a new approach to Quantify Novelty in Social Networks / Victor Maciel Guimarães dos Santos – Salvador, 2020.

51p.: il.

Orientador: Prof. Dr. Prof. Dr. Ricardo Araújo Rios.

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2020.

1. Grafos Temporais. 2. Concept Drift. 3. Redes Sociais. I. Rios, Ricardo Araújo. II. Universidade Federal da Bahia. Instituto de Matemática e Estatística. III Título.

CDD – XXX.XX
CDU – XXX.XX.XXX

TERMO DE APROVAÇÃO

VICTOR MACIEL GUIMARÃES DOS SANTOS

TEMPORAL NOVELTY QUANTIFICATION: A NEW APPROACH TO QUANTIFY NOVELTY IN SOCIAL NETWORKS

Esta Dissertação de Mestrado foi julgada adequada à obtenção do título de Mestre em Ciência da Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia.

Salvador, 30 de Julho de 2020

Prof. Dr. Ricardo Araújo Rios
Universidade Federal da Bahia

Prof. Dr. Angelo Conrado Loula
Universidade Estadual de Feira de Santana

Profa. Dra. Daniela Barreiro Claro
Universidade Federal da Bahia

ACKNOWLEDGEMENTS

I want to thank my friends and girlfriend for the company, advice, and motivation offered during the development of this project. While also acknowledging my family's support, especially my mother's, provided during this journey. To conclude, I express my gratitude to my advisor for his patience, collaboration, and contribution leading to the construction, evolution, and refinement of this project.

Stay awhile and listen...

—DECKARD CAIN

RESUMO

Com o aumento na adoção e uso das Redes Sociais, há um maior volume de dados online produzidos por usuários, que podem ser empregados para analisar seu comportamento nessas plataformas. Estas analises são úteis para, por exemplo, partidos políticos e empresas privadas, permitindo-lhes examinar como os usuários reagem a novas informações. Técnicas das áreas de Teoria dos Grafos, Séries Temporais e Aprendizado de Máquina estão entre as usadas para analisar o comportamento dos usuários. Ainda assim, essas técnicas produzem resultados melhores quando modelam certos aspectos desse comportamento, como a relação temporal ou a dependência entre termos utilizados nas publicações dos usuários. Este projeto considerou esses aspectos e hipotetizou que a adoção de grafos temporais, em conjunto com técnicas de Mineração de Texto e Series Temporais, permitem detectar mudanças de comportamento em usuários de Redes Sociais. Assim, para validar esta hipótese, foi desenvolvida uma nova abordagem que identifica pontos de mudança no comportamento dos usuários. Este procedimento utiliza técnicas de Mineração de Texto para encontrar termos, que serão utilizados posteriormente na criação de grafos temporais, mantendo seus relacionamentos nos textos originais e suas dependencias temporais. Em seguida, uma nova medida (*Temporal Novelty Quantification*), desenvolvida neste trabalho, é aplicada para quantificar como as opiniões dos usuários mudam com o tempo por meio de variações nas palavras usadas por eles. A utilização dessa medida em pares sequenciais de janelas de tempo gera uma série temporal que modela o comportamento dos usuários para um período observado. Finalmente, foi apresentado um método para detectar automaticamente mudanças de comportamento, visando identificar os pontos em que ocorrem estas mudanças. Além da abordagem apresentada, este trabalho contém um estudo de caso com sua utilização, a partir de um evento histórico no Brasil: as eleições presidenciais de 2018, que tiveram volume expressivo de publicações e efetivamente estabeleceram as Redes Sociais como o principal mecanismo para publicidade e ativismo político. Os resultados obtidos destacam eventos relevantes ocorridos na corrida presidencial, que podem ter levado a mudanças de comportamento nos usuários, mostrando o valor da abordagem desenvolvida. Este resultado também introduz novas possibilidades de pesquisa com base neste trabalho como, por exemplo, a identificação de *bots* que propagam *fake news*.

Palavras-chave: Grafo Temporal, Concept Drift, Redes Sociais

ABSTRACT

With the increase in adoption and usage of Social Networks, there is an uptick in the volume of user data produced online, which can be employed to analyze their behavior on these platforms. This information is useful, for example, to political parties and private companies, allowing them to examine how users react to new content. Techniques from the Graph Theory, Time Series, and Machine Learning fields are among the ones used to analyze user behavior. Still, methods from these fields provide better results by modeling certain aspects, such as the temporal relationship or dependency between terms present in user's publications. This project considered these aspects and hypothesized that the adoption of temporal graphs, in conjunction with techniques from the Text Mining and Time Series fields, allows the detection of changes in behavior on users of Social Networks. Thus, to validate this hypothesis, a new approach was developed that identifies changing points in users' behavior. This procedure uses Text Mining techniques to find terms, which will be used later in the creation of temporal graphs, maintaining their relationships in the original texts and their temporal dependencies. Then, a new measure (Temporal Novelty Quantification), developed in this work, is applied to quantify how users' opinions mutate over time through variations in words used by them. The utilization of this measure in sequential pairs of time windows generates a time series that models users' behavior for an observed period. Finally, a method for automatic detection of behavior change was presented, which aims to identify points when these changes occur. Besides the designed approach, this work also contains a case study for it, based on a remarkable event in Brazil: the 2018 presidential elections, which had an abnormal volume of publications and effectively established Social Networks as the leading mechanism for political activism and advertising. The results obtained highlight relevant events that happened in the presidential race, which could lead to changes in behavior for the users, and shows the value of the designed approach. This outcome also introduces new research possibilities based on this work, such as the identification of bots that propagate fake news.

Keywords: Temporal Graph, Concept Drift, Social Networks

CONTENTS

Chapter 1—Introduction 1

1.1 Context and Motivation 1
1.2 Hypothesis and Objective 3
1.3 Organization of Work 3

Chapter 2—Background 5

2.1 Graphs 5
2.2 Temporal Graphs 7
2.3 Time Series 10
2.3.1 Concept Drift 12
2.4 Related Work 13
2.5 Final Remarks 14

Chapter 3—Designed Approach 15

3.1 Initial Remarks 15
3.2 Phase 1: Data Collection 15
3.3 Phase 2: Preprocessing 15
3.4 Phase 3: Temporal Graph Construction 18
3.5 Phase 4: Time Series Extraction and Analysis 20
3.5.1 Extraction with Temporal Novelty Quantification 20
3.5.2 Analysis with Concept Drift 23
3.6 Final Remarks 25

Chapter 4—Case Study: Brazilian Politics in 2018 **election** 27

4.1 Initial Remarks 27
4.1.1 Experimental Setup 27
4.2 Data Collection Phase (Task 1) 28
4.3 Preprocessing Phase (Task 2) 28
4.3.1 URL Purge and Case Folding 30
4.3.2 Lemmatization 31
4.3.3 Tokenization 32
4.3.4 Stopwords Removal 32
4.3.5 Stemmization 32
4.4 Temporal Graph Construction (Tasks 3, 4 and 5) 33

4.5 Time Series Extraction and Analysis (Tasks 6 and 7) 34
4.6 Results 35
4.7 Final Remarks 37

Chapter 5—Conclusion 39

5.1 Discussions about our approach 39
5.2 Future work 40

LIST OF FIGURES

2.1 Simple graph. 5
2.2 Weighted digraph. 6
2.3 Weighted graph (G) with $rad\,(G) = 5$ and $diam\,(G) = 9$. 7
2.4 Simple directed temporal graph. 8
2.5 Time Varying Graph. 9
2.6 Edges exchanging weights in a Time Varying Graph. 10
2.7 Time series and its components. 12

3.1 Summary of our approach. 16
3.2 Example of every preprocessing step applied document-wise. 17
3.3 Temporal graph from Phase 3, Tasks 4 (a) and 5 (b), of our approach. . . 19
3.4 Temporal graph on instants P_1 (a) and P_2 (b). 20
3.5 Another temporal graph on instants P_3 (a) and P_4 (b). 20
3.6 Time series generated from Temporal Novelty Quantification data on Table 3.4. 23
3.7 Steps to identify Concept Drift points: Extrema identification (a), Interval definition (b) and Extrema selection (c). 24

4.1 Volume of published tweets related to Bolsonaro (a) and Lula (b). 29
4.2 Example of every preprocessing step applied document-wise. 30
4.3 Time series produced after analyzing tweets with reference to Bolsonaro. 34
4.4 Time series produced after analyzing tweets with reference to Lula. . . . 35

5.1 Volume of cloned tweets, without retweets, from Bolsonaro (a) and Lula (b). 41

LIST OF TABLES

2.1 Adjacency matrix for Figure 2.1. 6
2.2 Adjacency matrix for Figure 2.2. 6

3.1 Adjacency matrix for the temporal graph in Figure 3.3. 19
3.2 Topological Overlap for each vertex of the graphs from Figures 3.4 and 3.5 21
3.3 Distances for each edge of the graphs from Figures 3.4 and 3.5. 22
3.4 Temporal Novelty Quantification (TNQ) ($\aleph$) for each temporal graph in Figures 3.4 and 3.5. 22

4.1 Amount of tweets before and after preprocessing of the corpus. 28
4.2 Daily average of removed URLs and word accents for Bolsonaro and Lula. 31
4.3 Daily average of unique words in each preprocessing step for each politician. 33
4.4 Amount of vertices and edges for a temporal graph between days 1 (G_1) and 2 (G_2) for Lula and Bolsonaro. 33
4.5 Concept Drifts found for Bolsonaro and Lula. 35

5.1 Average and highest volume (27th) of duplicated tweets for both politicians. 40

NOTATION

(v, u)	edge of a graph
Δ	time interval
ϵ	graph eccentricity
μ	mean
ω	weight of an edge
π	pi value
ρ	presence function of a TVG
σ	standard deviation
τ	time interval considered on a TVG
Θm	time interval of a time series
θ	weigth function of an edge in a time instant
ε	random component of a time series
$\widehat{G}$	graph sequence
ζ	latency function of a TVG
A	adjacency matrix
a	cell in a graph adjacency matrix
C	topological overlap
d	distance between vertices
$diam$	graph's diameter
E	edge set of the graph
f	number of observed variables
G	graph

H	tendency
h	number of periods
i	index variable
j	index variable
k	time instant
M	amount of time instants in the temporal graph
m	time instant
n	amount of tweets
P	period chosen to be analyzed
p	sample word
Q	solution vertices for the P-Median problem
q	amount of vertices
r	amount of graphs
rad	graph radius
S	seasonality
s	amount of words
T	set of tweets
t	tweet in a set
u	graph vertex
V	graph vertex set
v	graph vertex
W	vector of words used in the temporal graphs
w	word in a tweet
X	set of observations
x	observation in a time instant
Y	list of points

y	point on the time series
Z	concept drift points
z	concept drift point

ACRONYM LIST

TVG	Time Varying Graph	8
NCD	Normalized Compression Distance	14
CRQA	Cross-Recurrence Quantification Analysis	14
URL	Uniform Resource Locator	16
NILC	Interinstitutional Nucleus of Computational Linguistics	31
NLTK	Natural Language Toolkit	32
TNQ	Temporal Novelty Quantification	39
BRACIS	Brazilian Conference on Intelligent Systems	39
USP	University of São Paulo	40
TF-IDF	Term Frequency–Inverse Document Frequency	40
FAPESP	São Paulo Research Foundation	40
AI	Artificial Intelligence	

Chapter 1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Historically, information and news were mainly spread by traditional media, such as well-known journalist programs broadcasted on TV, newspapers, radio stations, and magazines. The advent of Social Networks, streaming services, and microblogs have reduced this monopoly by allowing users to create content and information by themselves in an accessible manner, increasing the amount of content available. In this sense, Zwolenski and Weatherill (2014) stated that the annual volume of information available online would increase from 4.4 to 44 zettabytes (trillions of gigabytes) in 2020.

With their evolution, Social Networks have become part of people's lives, as evidenced in surveys made by the Pew Research Center (2018), Pew Research Center (2019), where 71 % of young Americans utilize them, at least once a day. When focused on users from Brazil, a report by Hootsuite and We Are Social (2020) states that in 2020, 66 % of the Brazilian population uses social media for three hours and thirty-one minutes per day, on average. Thus, we see that users from these platforms are constantly accessing and consuming digital materials distributed on them.

A decisive aspect of these platforms is that important information, educational or not, is easily propagated to a large number of users in a short time (TIRYAKIOGLU; ERZURUM, 2011). While this feature is applied to spread fake news and promote bullying (ZHAO; ZHOU; MAO, 2016), it can also lead to the creation of movements that give and amplifies the voices of users, like "Black Lives Matter" [1], "#MeToo" [2], and "#BrequeDosAPPs" [3]. Such movements are entirely conceived inside Social Networks, relying on active consumption of content by social media users for their effectiveness.

[1] See: ⟨https://blacklivesmatter.com/⟩

[2] See: ⟨https://g1.globo.com/pop-arte/games/noticia/2020/06/22/me-too-dos-gamers-leva-mulheres-a-relatarem-casos-de-abuso-na-industria-dos-jogos.ghtml⟩

[3] See: ⟨https://www.diariodepernambuco.com.br/noticia/brasil/2020/07/breque-dos-apps-greve-dos-entregadores-acontece-nesta-quarta.html⟩

The large-scale production of content published on Social Networks, as a consequence of their growth and adoption, takes the form of photos, videos, and texts. Generally, these publications reflect users' personal opinions and behaviors since they are the means of expression available on these platforms and have great value to companies and other interested parties, such as political groups. Companies might be interested in mining such opinions, for example, to monitor the satisfaction of its customers or to identify new potential clients for future advertising campaigns. Meanwhile, the interest of political parties would be in analyzing voters' feelings on important issues, such as pension reforms or internal affairs, and also to monitor the spread and impacts of fake news.

Another consequence of such platform growth and adoption is the motivation of studies based on them, such as their usage as a transnational space to monitor migrants (LUBBERS; VERDERY; MOLINA, 2020) and as tools for discrimination in hiring processes (ACQUISTI; FONG, 2020). Meanwhile, other research focuses more on the content produced by users either to extract information from multimedia content (HERMIDA; HERNÁNDEZ-SANTAOLALLA, 2018), perform analysis of textual publications (RIOS et al., 2017a; MELLO et al., 2018), comprehend disease diffusion (ZHANG; CENTOLA, 2019), or model rumor spreading (YANG; LI; GIUA, 2020).

Computational approaches allow the analysis of information from Social Networks and, therefore, the development of new studies. In general, approaches based on Graph Theory have nodes representing data, such as users, which are connected by edges that describe relationships between them (ZACHARY, 1977; ZHANG et al., 2010; ZHU et al., 2016; CAO et al., 2020; BREUER; EILAT; WEINSBERG, 2020). An alternative form to model these platforms information can be seen in (SUBBIAN; PRAKASH; ADAMIC, 2017; RIOS et al., 2017a; RIOS et al., 2017b; MELLO et al., 2018; LI et al., 2019), where the texts extracted from social networks are transformed into time series to investigate temporal variations. Another possible way to represent such data is seen in (LUO et al., 2019), which transforms then into an attribute-value matrix for later use with machine-learning algorithms.

However, each method has its drawback when representing data obtained from Social Networks. For example, the time series approach might fail to include the relationships between different variables, whereas graphs do not usually consider temporal aspects. The latter also occurs when using an attribute-value matrix, which does not consider changes in variable relationships over time, further hindering the transformation of these data for usage with Machine Learning tools.

To tackle this problem, we design a new challenging approach by combining Machine Learning, Graph Theory, and Time Series Analysis to better deal with Social Network data, seeking to maintain the temporal relationship of terms present in user's publications. We assessed our approach on texts published on Twitter (tweets), collected during the 2018 Brazilian presidential election race, which was characterized by intense participation of politicians, political parties, voters, and bots. In Brazil, the election was a milestone in history due to the amount of published content, its political relevance, and by changing the way people do political advertisement and activism.

1.2 HYPOTHESIS AND OBJECTIVE

In this work, we developed a new approach to handle the problem of modeling information from Social Networks and its relationships while maintaining changes that occur over time. This approach also allows us to analyze users' behavior through variations of words used in a time period. As such, we have defined the following hypothesis to guide our research:

> *"The adoption of Temporal Graphs along with Preprocessing and concept drift methods for Time Series allows the detection of changes in users' behavior on Social Networks, related to specific topics, through shifts in the volume of co-occurrences between words used by them."*

We planned this approach in four Phases: 1) textual data related to specific topics of interest published on Twitter were collected and stored; 2) application of Text Mining methods to extract relevant words employed to express users' opinions; 3) usage of these words to create temporal graphs that represent their connections and temporal relationships; and 4) utilization of a measure proposed in this work to transform the temporal graph into a time series, which was, then, ready to be further analyzed. Lastly, in the context of this work, we also proposed a new straightforward method to detect concept drifts, i.e., the detection of changing points in people's behavior, which can be induced by some real-world events.

Each of the last three Phases also constitutes specific objectives of this research aiming to refine the textual data obtained (Phase 2), build temporal graphs from words (Phase 3), extract a time series from temporal graphs, and perform analysis on it (Phase 4). Meanwhile, the main goal of this dissertation was the validation of the hypothesis presented previously. To summarize, this research present the following contributions to the state of the art:

- A new approach to model published texts into structures (graphs and time series) that preserve their temporal relationships;
- A new distance measure (Temporal Novelty Quantification - $\aleph$) to convert temporal graphs into time series;
- A new method to detect changes in users' behavior.

Finally, we evaluated the importance of our approach by associating the detected changes in behavior (novelties) with news registered by traditional media.

1.3 ORGANIZATION OF WORK

This dissertation is organized as follows: Chapter 2 introduces the main concepts used in this work and related researches; the details of our approach are in Chapter 3; Chapter 4 explains one study case and review the obtained results; Finally, concluding remarks and future directions are present in Chapter 5, which is followed by a list of references.

Chapter 2

BACKGROUND

2.1 GRAPHS

Formally, a graph G is defined as a finite and non-empty set $G = \langle V, E \rangle$, in which $V(G)$ represents a set of elements, called vertices or nodes, and $E(G)$ symbolizes a set of distinct and unordered pair of elements from $V(G)$, called edges (CORMEN et al., 2009; WILSON, 1996). An edge $\{v, u\}$ is often said to connect vertices v and u with its abbreviated form being (v, u) (CORMEN et al., 2009). Figure 2.1 illustrates a graph with $V(G) = \{v, u, c\}$ and $E(G) = \{(v, u), (v, c), (u, c)\}$.

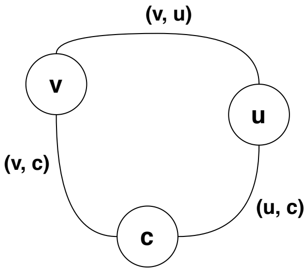


Figure 2.1: Simple graph.

Another way to represent a graph is by using an adjacency matrix, in which Equation 2.1 (CORMEN et al., 2009) is used to depict all edges. Table 2.1 presents the adjacency matrix for the graph in Figure 2.1.

$$a_{(v,u)} = \begin{cases} 1 & \text{if } (v, u) \text{ or } (u, v) \in E\,, \\ 0 & \quad\quad \text{otherwise}\,. \end{cases} \tag{2.1}$$

Graph nodes have been widely adopted to model countless real-world problems, such as: i) cities (APPLEGATE et al., 2007); ii) resources (HAKIMI, 1964; HAKIMI, 1965); iii) people (BALBONI et al., 2015); and iv) words (KATRAGADDA et al., 2016). Meanwhile, edges also have been applied to describe several relationships. For

Vertex	v	u	c
v	0	1	1
u	1	0	1
c	1	1	0

Table 2.1: Adjacency matrix for Figure 2.1.

example, word co-occurrences (KATRAGADDA et al., 2016), and the connection between places (APPLEGATE et al., 2007; HAKIMI, 1964; HAKIMI, 1965) and people (BALBONI et al., 2015).

Graph edges may present associated weights or costs (ω), thus indicating a value spent to move between nodes, e.g., the distances between cities or countries. Graphs whose edges have weights are usually called weighted graphs (WILSON, 1996).

Edges can also contain directions, which means that the relationship defined by them occurs only in one way. In this case, ordered pairs of connected vertices called arcs describe the edges of the graph (GODSIL; ROYLE, 2001). Graphs with edges that contain such information are called directed graphs or digraphs. In Figure 2.2, there is a weighted digraph whose edges $\{(v,u)$, (u,c), (c,v), $(v,c)\}$ have, respectively, the weights {3, 2, 6, 4}.

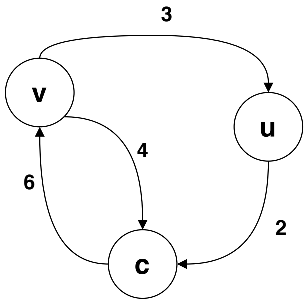


Figure 2.2: Weighted digraph.

Note that edges (c,v) and (v,c) include the same vertices, however, their weights and directions are different from each other. It is also observable that there is neither a direct link between u and v nor between c and u. Table 2.2 also depicts this behavior, showing the adjacency matrix for the weighted digraph in Figure 2.2 using Equation 2.2 (CORMEN et al., 2009).

Vertex	v	u	c
v	0	3	4
u	0	0	2
c	6	0	0

Table 2.2: Adjacency matrix for Figure 2.2.

$$a_{(v,u)} = \begin{cases} \omega(v,u) & \text{if } (v,u) \in E\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.2}$$

Another important concept for graphs is a path, which is a set of unique vertices, sequentially and directly connected through an edge. A set of nodes $\{v_1, v_2, v_3, \ldots, v_{q-1}, v_q\}$ forms a path if and only if there is an edge (v_i, v_{i+1}) and v_{i+1} immediately follows v_i in the node-set.[1] The distance between vertices v and u, represented as $d(v,u)$, is the minimum amount of edges in a path from v to u. Meanwhile, for weighted graphs, it corresponds to the minimum sum of edge weights in a path going to u from v (WILSON, 1996; WEST, 2017).

One may notice that by using concepts, such as distance, it is possible to extract objective information from graphs. These concepts are known as graph metrics and measures, with other examples being the eccentricity, radius, and diameter (WEST, 2017).

The eccentricity $\epsilon(v)$ of a vertex v from graph G corresponds to the largest distance from v to any other vertex of G, i.e., $\epsilon(v) = max_{\, u \in V(G)}\, d(v,u)$. Meanwhile, the radius $rad\,(G)$ and diameter $diam\,(G)$ of a graph match, respectively, with the lesser and greater eccentricity of all vertices, such that $rad\,(G) = min_{\, v \in V(G)}\, \epsilon(v)$ and $diam\,(G) = max_{\, v \in V(G)}\, \epsilon(v)$ (WEST, 2017). In Figure 2.3, vertex v of a weighted graph G has $\epsilon(v) = 5$ considering a path from v to c passing through u, which is also the graph radius $(rad\,(G) = 5)$. Meanwhile, its diameter is equal to 9, found while going from c to b, passing by u and v.

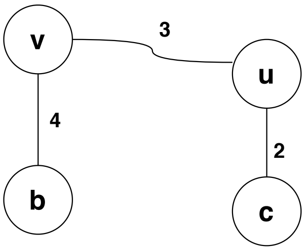


Figure 2.3: Weighted graph (G) with $rad\,(G) = 5$ and $diam\,(G) = 9$.

Despite the versatility previously discussed on graphs, their lack of mechanisms to deal with temporal information, which leads them to be called static graphs, has motivated the design of a new research branch referred to as temporal graphs (WANG et al., 2019), presented in the following section.

2.2 TEMPORAL GRAPHS

Temporal graphs are an extension of the definitions previously presented with the inclusion of temporal information to model networked time-evolving systems (PAN; SARAMÄKI, 2011). The flexibility to represent content from such systems makes the use of temporal graphs an object of interest in several research domains such as Computer Science, Medicine, and Economics.

[1] In this work,"i" and "j" are just iteration variables used in different contexts.

Computer Science has used temporal graphs to perform a literature review (ERTEN et al., 2004); obtain a summary of videos and identify their scenes (NGO; MA; ZHANG, 2005); and detect human movements in Youtube videos (AOUN; MEJDOUB; AMAR, 2014). In Medicine, their usage is seen, for example, to diagnose epilepsy (TANG et al., 2013) and to check the brain dynamics in patients with schizophrenia (YU et al., 2015). Finally, in Economics, we can exemplify their application with the research presented by Stephen, Gu and Yang (2015), which makes predictions about the behavior of the North American stock exchange.

One form of modeling using temporal graphs is to construct a single aggregated static graph, in which all connections between each pair of nodes over time are flattened into a single edge. Such representation is relevant to investigate the structure and function of systems in which the topological characteristics are more significant than the temporal ones (NICOSIA et al., 2013). Figure 2.4 shows an example of this representation.

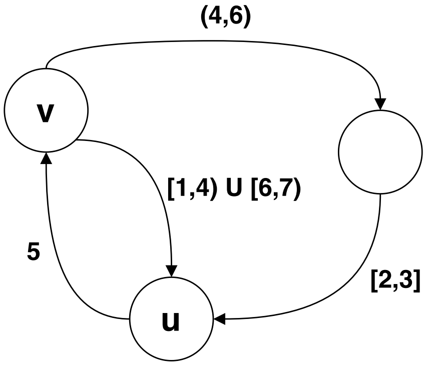


Figure 2.4: Simple directed temporal graph.

The numbers associated with each edge indicate the interval in which it was present. These intervals follow the mathematical set notation, meaning that the edge between v and u, for example, occurs only at moments $\{1, 2, 3, 6\}$. Although this representation allows to include the time interval in which two vertices are connected, it is not flexible enough to depict systems in which temporal aspects are relevant (NICOSIA et al., 2013), for example, when there are multiple intervals of connections over time. Such a limitation has motivated the development of Time Varying Graphs.

According to Casteigts et al. (2012), a Time Varying Graph (TVG), G, is defined as $G = \langle V, E, \tau, \rho, \zeta \rangle$. In this sense, V and E are equivalent to the definition of static graphs, thus representing the sets of vertices (V) and edges (E) of the temporal graph, noting that vertices, edges, and edge directions might change over time. $\tau \subseteq \mathbb{T}$ indicates the time interval considered by the graph, with the domain $\mathbb{T}$ usually being $\mathbb{R}$ for continuous-time systems or $\mathbb{N}$ for discrete-time systems, whereas ρ and ζ represent functions of the graph.

The function $\rho : E \times \tau \to \{0, 1\}$, called presence function, indicates the existence of an edge in a given moment. The latency function, defined as $\zeta : E \times \tau \to \mathbb{T}$, shows the amount of time it takes to traverse an edge, starting in a specific time instant (CASTEIGTS et al., 2012).

Then, a TVG, or temporal graph from this point forward, may be described as a sequence of static graphs $G = \{G_1, G_2, \ldots, G_r\}$, where each element G_i of the sequence represent the state of the temporal graph in a time instant $m \in \tau$ (NICOSIA et al., 2013), as illustrated by Figure 2.5. In this figure, one may notice the behavior of the temporal graph as a timeline of edge occurrences.

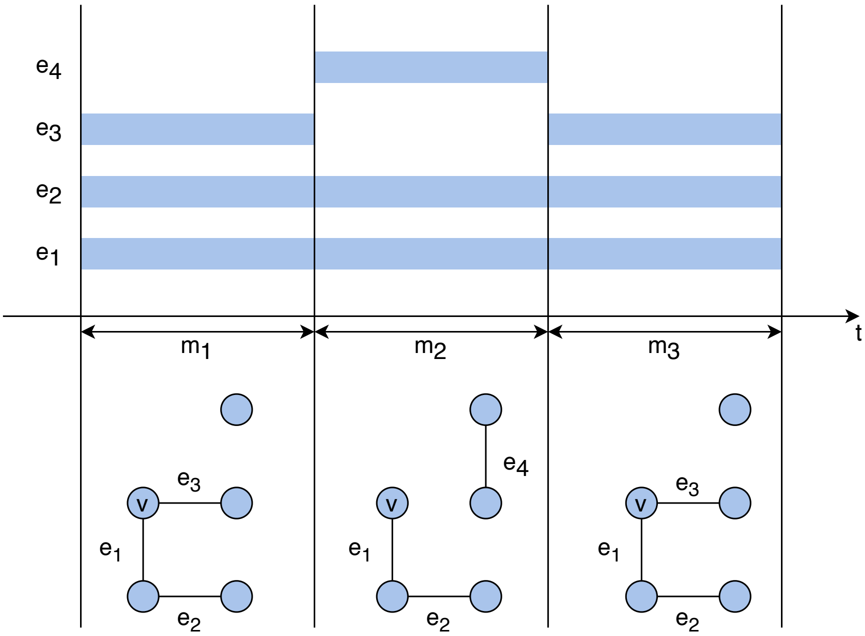


Figure 2.5: Time Varying Graph.

The example in Figure 2.5 is composed of three time windows (m_1, m_2, and m_3) and four edges (e_1, e_2, e_3, and e_4) connecting five vertices. By looking at this representation, we notice edge e_3 models a relationship between vertices during two different time windows. Moreover, e_4 only shows up during a single window, while edges e_1 and e_2 persist during the whole analysis. The identification of edge variations between windows is one of the goals of this work.

After modeling a system with TVGs, it is possible to extract and quantify information from them, with the application of metrics and measures, such as the temporal distance and diameter (NICOSIA et al., 2013). Another well-known measure is Topological Overlap (NICOSIA et al., 2013; CLAUSET; EAGLE, 2007), described in Equation 2.3, which aims at evaluating the persistence of edges from a node on the temporal graph between two consecutive time intervals.

$$C_v(m, m+1) = \frac{\sum_u a_{(v,u)}(m) a_{(v,u)}(m+1)}{\sqrt{\left[\sum_u a_{(v,u)}(m)\right] \left[\sum_u a_{(v,u)}(m+1)\right]}} \tag{2.3}$$

At this equation m and $m+1$ refer to the time windows to be evaluated, v and u represent nodes of the unweighted temporal graph, and $a_{(v,u)}(m)$ is the entry in the graph adjacency matrix for v and u at time m, representing if the edge between then is present on moment m. In case $C_v(m, m+1) = 1$, there are no changes in the edges of v between instants m and $m+1$. On the other hand, if $C_v(m, m+1) = 0$, then no edge of v at instant m is present in $m+1$ and vice versa.

From this measure, it is possible to extract the average topological overlap of a vertex based on the nodes connected to it. In summary, this measure is the mean of topological overlaps obtained for the vertex during the whole period of the temporal graph (NICOSIA et al., 2013; TANG et al., 2010b). Equation 2.4 shows how to obtain this mean value.

$$C_v = \frac{1}{M-1} \sum_{m=1}^{M-1} C_v(m, m+1) \tag{2.4}$$

In this equation, M is the whole analyzed interval. In essence, this value estimates the probability of edges to persist between two consecutive time windows and also captures the tendency of these edges to persevere across multiple windows (NICOSIA et al., 2013).

Thus, by going back to Figure 2.5 and considering edge e_3, the only time when it does not occur is m_2. Then, the topological overlap of v, between moments m_1 and m_2 is 0.7. The average topological overlap of v is also 0.7, since its topological overlap between m_2 and m_3 is again 0.7. However, in this work, we have noticed the following problems with these measures:

- The topological overlap does not take into account weighted temporal graphs and the weights of its edges;
- Even if the topological overlap considered edge weights, it would still hide particular changes on temporal graphs, for example, when edges from a vertex trade weights between them, as seen in Figure 2.6.

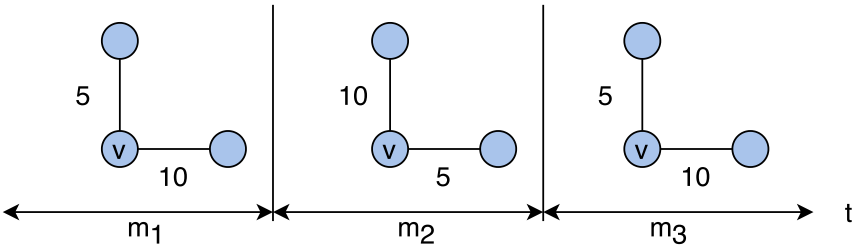


Figure 2.6: Edges exchanging weights in a Time Varying Graph.

These issues led us to create a new measure, presented in Section 3.5.1. Regardless of the one adopted, by comparing two consecutive time instants, we can monitor how the system is evolving through the organization of the measure outputs as a time series, whose main concepts are detailed in the next section.

2.3 TIME SERIES

Time series are usually adopted to organize a set of observations collected over a given time interval $X_n = \{x_0, x_1, x_2, \ldots, x_m\}$, where x_0 is an observation collected at the initial instant 0 and x_m is another observation collected at the time instant m (BOX et al., 2015; MORETTIN; TOLOI, 2006; SHUMWAY; STOFFER, 2006; CHATFIELD, 2004; RIOS, 2013).

The importance of employing time series to analyze systems is noticed in several areas of study as, for example, Economy, Biology, Medicine, and Weather. In Economics, there exist several studies (GUHATHAKURTA; BHATTACHARYA; CHOWDHURY, 2010; LEBARON; ARTHUR; PALMER, 1999; CHOI; HAUSER; KOPECKY, 1999), which aim to model fluctuations and predict values in stock markets.

In Biology and Medicine, different researches use time series to model symptoms and evolution of diseases, tendencies for the emergence of cancer, studies of heartbeats or brain signals, and others (TSCHACHER; KUPPER, 2002; HOSOKAWA et al., 2003; SUMMA et al., 2007; ZHUANG et al., 2008; PONOMARENKO et al., 2005; RAIESDANA et al., 2009; CRABTREE et al., 1990; JUANG et al., 2017).

Similarly, the usage of time series has attracted the attention of researchers to model the behavior of the planet climate change (KOÇAK; SAYLAN; EITZINGER, 2004; YU; CLARK; LEONARD, 2008; KÄRNER, 2009; KO et al., 1993; FEARNSIDE, 1999; KUMAR; JHA, 2013)

According to Box et al. (2015), time series can be classified into univariate or multivariate. Univariate time series are composed of scalar values sequentially collected over time using the form $\{x_1, x_2, \ldots, x_m\}$. However, when f variables ($f > 1$) are observed during each time instant m, the time series is said to be multivariate, denoted by $\{x_{1m}, x_{2m}, \ldots, x_{fm}\}$ (BOX et al., 2015).

Regarding the interval between collections, time series are subdivided into two classes (MORETTIN; TOLOI, 2006): I) Discrete, whose analysis is made on the temporal domain, according to time intervals Θm both fixed and periodic in $\mathbb{N}$; and II) Continuous, whose analysis is made on the frequency domain (time in $\mathbb{R}+$).

It is important to highlight that every observation is defined by a combination of different influences. As discussed by Morettin and Toloi (2006) and Box et al. (2015), a given time series X_m can be denoted by the sum of three non-observable components $X_m = H_m + S_m + \varepsilon_m$, in which H_m represents a tendency, S_m a seasonality and ε_m a random component (MORETTIN; TOLOI, 2006). H_m describes variations on the behavior of the series, and S_m indicates whether a particular behavior of the time series tends to repeat itself in a Θm time interval. These components, $\{H_m, S_m, \varepsilon_m\}$, are called not observable as they are not gathered directly from a system, being presumed through temporal relationships between observations.

Figure 2.7(a) presents an example of a time series that combines those three components, similarly as collected from real systems. Figure 2.7(b) exemplifies a seasonal behavior, which was created from a sinusoidal component with an angular frequency equal to π. Finally, Figures 2.7(c) and (d) present a positive tendency, and a noise created from a normal distribution (with mean equal to 0 and standard deviation equal to 0.5), respectively.

By understanding these components, global aspects of time series can be evaluated. Examples of such aspects are stochasticity, stationarity, and linearity. The correct understanding of these aspects is important to determine more accurate models for time

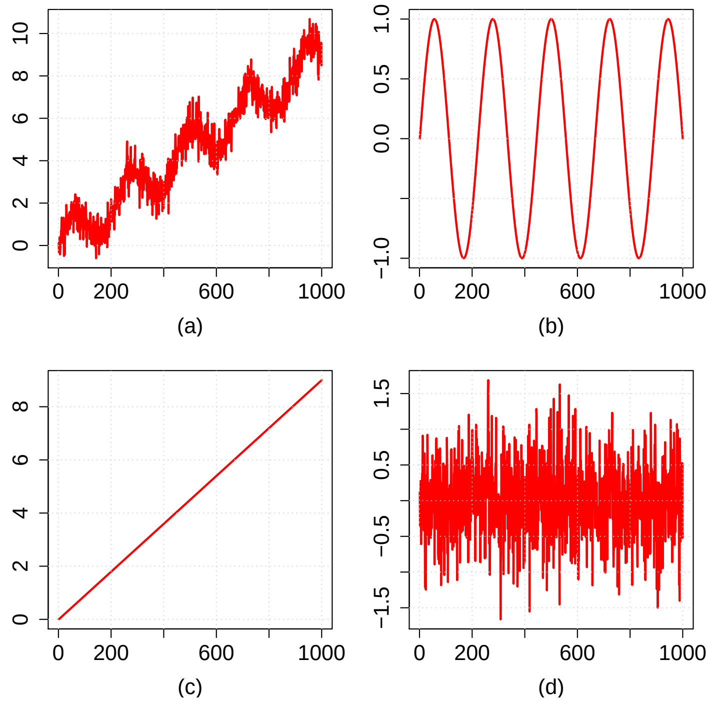


Figure 2.7: Time series and its components.

series.

Stochastic time series are made up of observations and random relationships that follow probability density functions and can change over time, making the modeling of their events difficult. In contrast, deterministic series predominantly presents observations with strict dependencies on past values. Stationary time series find themselves in a particular state of statistical balance (BOX et al., 2015), i.e., they develop randomly on time, around a constant average (MORETTIN; TOLOI, 2006). Linear time series are those whose observations are composed of a linear combination of past occurrences and noise. In turn, non-linear series are formed by a combination process of non-linear observations and past noises.

In the context of this work, outputs produced by our approach were organized as stochastic and non-linear time series in a univariate manner with discrete intervals. The resulting time series represents an observation of a variable, that can be further investigated using techniques from the Concept Drift field, which are discussed next.

2.3.1 Concept Drift

While observing variables of interest, they might have a dependency in some hidden or unknown context. As such, changes that occur in this context can affect the observed variable behavior and are commonly known as concept drifts (SCHLIMMER; GRANGER JR., 1986; WIDMER; KUBAT, 1996).

The detection of concept drifts is primarily related to online supervised learning (GAMA et al., 2014), with traditional algorithms such as STAGGER (SCHLIMMER; GRANGER JR., 1986) or FLORA (WIDMER; KUBAT, 1996) requiring (TSYMBAL, 2004; GAMA et al., 2014):

- a time window to detect possible changes;
- the use of labels for data classification.

However, the latter is prohibitive in our context as there is no label to describe the values in the series. Thus, a straightforward algorithm has been proposed to identify concept drifts on time series data, detailed in Section 3.5.2.

Since the produced time series model the behavior of users from Social Networks through texts published by them, the identified concept drift points indicate moments when these behaviors, about specific themes, change, which is the main objective of this research.

The next section presents related works available in the literature, whose main objective is to model the behavior of users in Social Networks using Graphs and Time Series concepts.

2.4 RELATED WORK

In addition to the introductory definitions presented so far, we also looked in the literature for researches that use similar approaches to model and analyze social network data.

In the first related study, the authors collected tweets containing a specific keyword while also referencing either of the two dominant presidential candidates and their respective political parties during the Spanish election of 2011 (BORONDO et al., 2012; BORONDO et al., 2016). From this data, analyzes were made using Time Series and Complex Networks concepts to better understand the behavior of politicians and voters during the election period. The time series analysis extracted the volume of tweets related to the candidates and political parties then used the Relative Support parameter, with results resembling the actual ratio of votes obtained by the two main parties. Proposed in (BORONDO et al., 2012), this parameter consists of a proportion between mentions' growth rates of two subjects.

Similarly, Caldarelli et al. (2014) collected tweets related to the main political parties, and their respective candidates, in the Italian election of 2013 and modeled the volume of data as time series. Next, the authors used the same Relative Support parameter to ratify the relationship between the election results and political activism on Twitter. However, these researches do not explore detecting changes in users' behavior or the relationship between words present on users' tweets.

In the literature, there is extensive material presenting different analyses of publications in Social Networks to understand real-world events. A straightforward way to perform this analysis is by using the volume of publications. Another area usually explored in such context is the analysis of sentiments, as seen in (RIOS et al., 2017a). This work collects data from Twitter related to the 2016 Brazilian impeachment regarding specific subjects. From it, time series are extracted, including one that summarizes the sentiment of tweets for each day and each subject. Finally, the peaks and spikes found in these series are analyzed and linked with real-world news and events. The main problem with this approach is the execution of black-box models, which conceal information about the words selected and their temporal relations.

The relationship between Social Networks and politics is also explored in (MELLO et al., 2018). It collected tweets linked to certain Brazilian politicians and transformed the volume of tweets into time series using Normalized Compression Distance (NCD) and sentiment analysis producers from the TSViz project (RIOS et al., 2017b). Then, on every resultant time series, the authors applied the algorithm Cross-Recurrence Quantification Analysis (CRQA) to detect changes in users' series behavior. The authors' goal was to identify such changes and connect them with events happening in the Brazilian political scenario. However, it does not explore the relation of words in the tweets.

Social Networks, such as Twitter, are powerful means of spreading information due to their short messages, the variety of information sources, and the large volume of data (KATRAGADDA et al., 2016). These characteristics hinder the application of traditional algorithms to detect and track topics, or events, in well-formatted data. Katragadda et al. (2016) proposed a new approach to detect events within the first 8 minutes of their occurrence.

To achieve this purpose, the authors: i) Tokenize and clean special characters, stop words, URLs, and retweet identifiers from all tweets collected in a time window; ii) Create a temporal graph, whose vertices are word tokens that appeared on the last m windows, and the edges indicate the number of tweets where the two vertices appear together; iii) Prune the graph, removing words not considered emergent or important, and then cluster it using a voltage based clustering algorithm; and iv) Filter obtained clusters, through comparisons against previous ones using the Kullback-Leibler divergence score, to extract only valid events. Although this approach analyzes the relationship between terms used on tweets, the change in users' behavior is not considered.

By taking into account the advantage of those researches, we present a new approach to transform a set of tweets published on a specific topic into a time-variant graph that preserves not only the relationship among their words but also their temporal dependencies, as detailed in the next chapter.

2.5 FINAL REMARKS

The Temporal Graph concepts, their respective metrics, and measures, along with Time Series definitions that have been presented so far are essential to a deeper understanding of this work. This chapter also introduced researches related to this work, which illustrates investigations made in the Social Network Analysis field. In the next chapter, we detail the modeling used in our approach to monitor users' behavior and to detect changes over time.

Chapter
3

DESIGNED APPROACH

3.1 INITIAL REMARKS

This chapter describes, in detail, the approach used in this research, developed to tackle the problem of modeling information from Social Networks and its relationships while maintaining changes that occur in them over time. Through it, we analyzed textual data representing users' behavior in Social Networks and indicated points where the public opinion might have differed on specific topics. Figure 3.1 shows the four phases of this approach.

3.2 PHASE 1: DATA COLLECTION

The first phase relates to the process of collecting data from a Social Network, as highlighted by Task 1 at the top-left plot of Figure 3.1. There are multiple methods to obtain this information, going from official APIs and tools to external software that collects data from these platforms, such as TSViz (RIOS et al., 2017b). How the data is collected or stored are beyond the scope of this research.

As a result of this phase, there is a multiset $T_\Delta = [t_1, t_2, t_3, \dots, t_n]$, in which Δ represents the monitored time interval and t_i, $1 \leq i \leq n$, is a given publication composed of w strings, which are sequences of alphanumerical characters, i.e., $t_i = \{w_1, w_2, \dots, w_s\}$.

3.3 PHASE 2: PREPROCESSING

The second phase wraps all preprocessing performed on the collected data, using some techniques that are widely adopted in the Text Mining field. The first change made in Task 2 was to remove all republished information once they are simply repetitions of other texts, and we are trying to identify novelty as people react to different topics. Thus, reducing the dataset to $T'_\Delta \subseteq T_\Delta$, that will be referred to as T_Δ from now on to simplify reading. Another significant transformation was the discretization of the observed period. Although we are monitoring a topic during time Δ, our focus is to understand the variations between time intervals within this period. Then, T_Δ is discretized considering h intervals of time

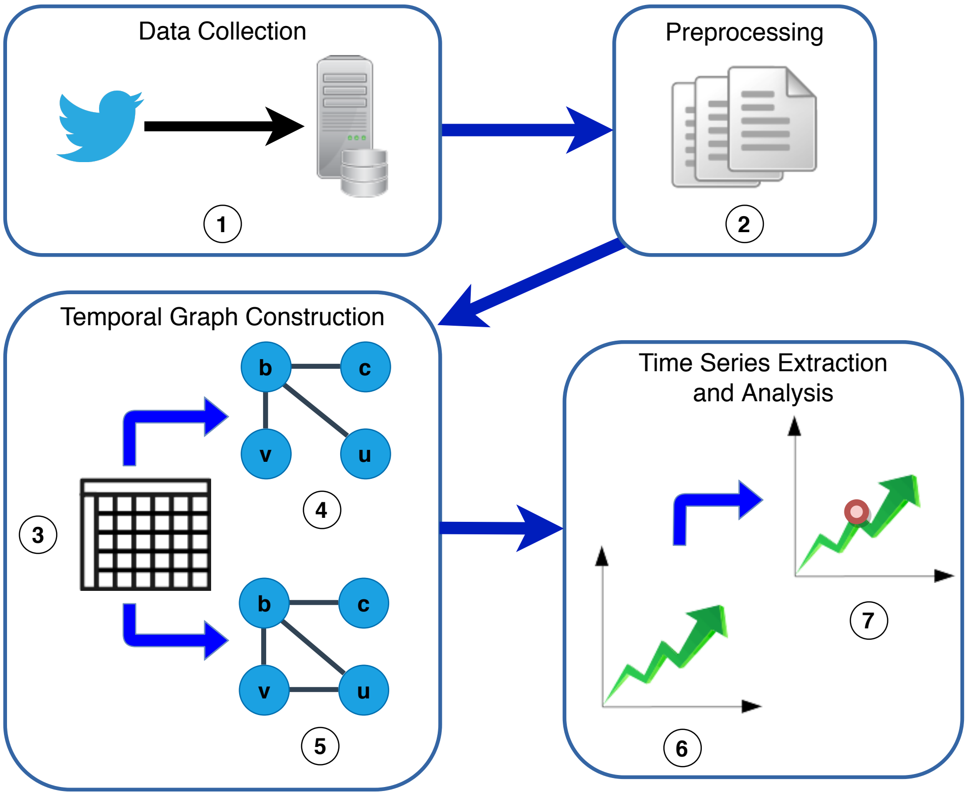

Data Collection
1
Preprocessing
2
Temporal Graph Construction
3
b
c
v
u
4
b
c
v
u
5
Time Series Extraction and Analysis
6
7

Figure 3.1: Summary of our approach.

with length P, thus producing a set of sequential intervals $T_\Delta = \{T_{P_1}, T_{P_2}, \dots, T_{P_h}\}$, in which $T_{P_i} = [t_1, \dots, t_n]$, T_{P_h} is the hth time window containing the last n publications, and $\Delta = \bigcup\{P_1, P_2, \dots, P_h\}$. For example, in this work, we have monitored some topics for 30 days, but we analyze them by using daily windows.

After these steps, the preprocessing starts modifying each window T_{P_i} and its publications. They go through URL purging, lemmatization, tokenization, removal of stopwords, and stemmization. Each of these transformations is detailed next, with Figure 3.2 summarizing and exemplifying them.

Uniform Resource Locator (URL) or link is a compact string representation for a resource available on the Internet (BERNERS-LEE; MASINTER; MCCAHILL, 1994). Considering the context of Social Networks, these links usually point to other media types or external content, meaning the actual content would be at the link destination. As such, in URL purge, the goal is to remove these links since they do not have any relevant information themselves. Thus every publication $t \in T_{P_i}$ could have part of its content removed, which results in a new $t' \subseteq t$, called t from now on.

When looking at texts, it is frequent to find multiple words that share the same grammatical root, such as verbs conjugated in different tenses or nouns being singular or plural. These variations would be considered distinct items, which is not helpful for this analysis. With lemmatization (MANNING; RAGHAVAN; SCHÜTZE, 2010), every publication t has its words reduced to their grammatical root, converting them into their base form. For example, the lemmatization of “went” produces “go”. Thus, a publication t go from $\{w_1, w_2, \ldots, w_s\}$ to $\{w'_1, w'_2, \ldots, w'_s\}$ where w'_i is the base form of w_i, $\forall\, w_i \in t$. For simplicity, from this moment forward, each word w'_i of t will be called w_i.

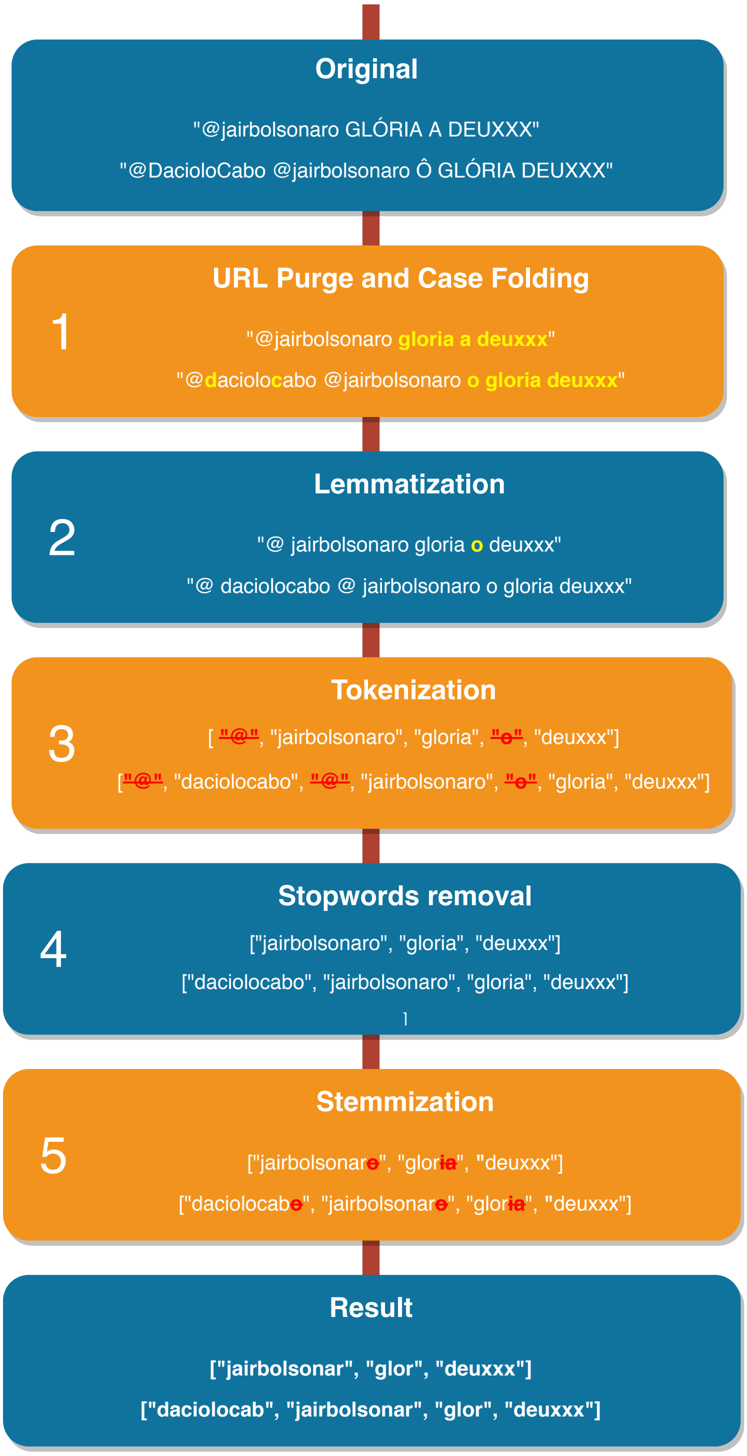

Original
"@jairbolsonaro GLÓRIA A DEUXXX"
"@DacioloCabo @jairbolsonaro Ô GLÓRIA DEUXXX"
1
URL Purge and Case Folding
"@jairbolsonaro gloria a deuxxx"
"@daciolocabo @jairbolsonaro o gloria deuxxx"
2
Lemmatization
"@ jairbolsonaro gloria o deuxxx"
"@ daciolocabo @ jairbolsonaro o gloria deuxxx"
3
Tokenization
[~~"@"~~, "jairbolsonaro", "gloria", ~~"o"~~, "deuxxx"]
[~~"@"~~, "daciolocabo", ~~"@"~~, "jairbolsonaro", ~~"o"~~, "gloria", "deuxxx"]
4
Stopwords removal
["jairbolsonaro", "gloria", "deuxxx"]
["daciolocabo", "jairbolsonaro", "gloria", "deuxxx"]
]
5
Stemmization
["jairbolsonar~~o~~", "glor~~ia~~", "deuxxx"]
["daciolocab~~o~~", "jairbolsonar~~o~~", "glor~~ia~~", "deuxxx"]
Result
["jairbolsonar", "glor", "deuxxx"]
["daciolocab", "jairbolsonar", "glor", "deuxxx"]

Figure 3.2: Example of every preprocessing step applied document-wise.

The identification of a text basic units, whether they are words, phrases, or another form, is the goal of tokenization (WEBSTER; KIT, 1992). With it, a publication t becomes a multiset of separated tokens (e.g., words, emojis, symbols), instead of a single block of text, meaning that publication t transforms from $\{w_1, w_2, \dots, w_s\}$ to $[w_1, w_2, \dots, w_s]$. This process is relevant since it outputs the textual information in a format that can be used in the next steps.

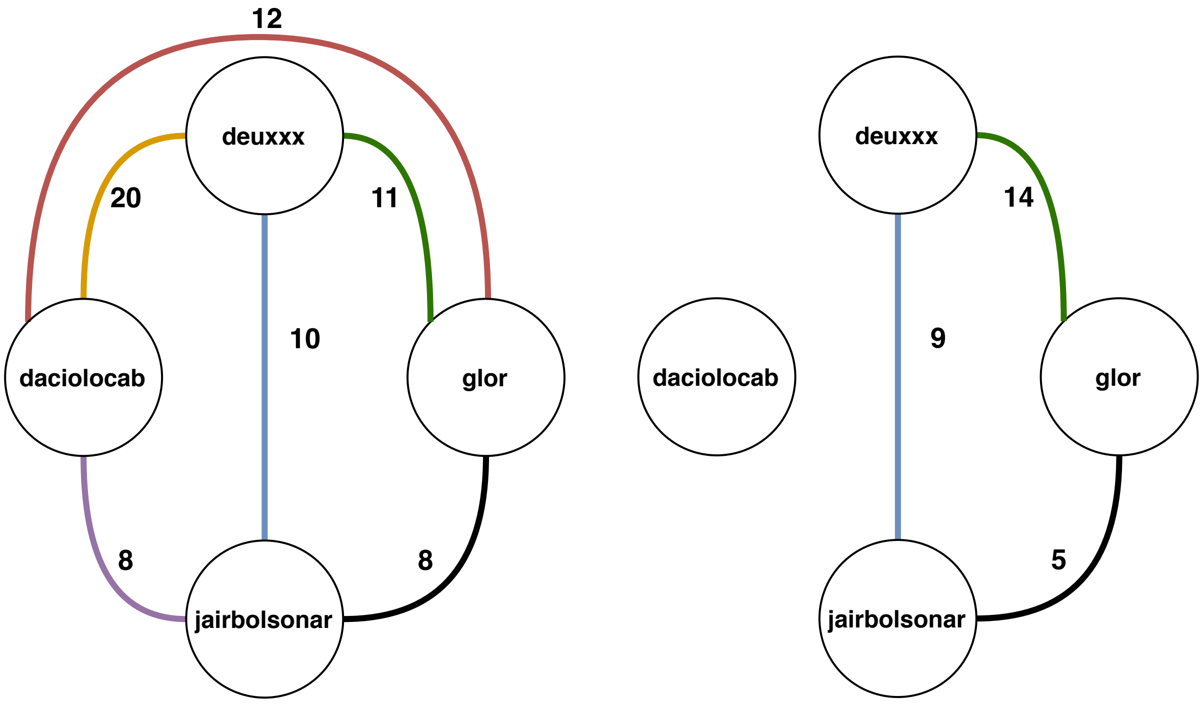
The subsequent transformation will remove stopwords, defined by Luhn (1960) as non-significant or frequently repeated words and symbols, e.g., prepositions, conjunctions, and articles. Some of these terms appear in lists previously mapped for most languages, with each analysis having the possibility of incrementing this list with terms specific to it, such as retweet ("RT") for Twitter data. However, this is not useful to our analysis since any republished information (retweet) would be already removed by then. Regardless, we removed stopwords from the publications since they do not bring new information to our analysis, meaning that words related to t will decrease to a new t' such that $t' = [w_i, w_j, \dots, w_p] \subseteq t$, which shall be called t from this point forward.

Finally, we applied stemmization (LOVINS, 1968) to each term in a publication. It has goals similar to lemmatization, usually performing straightforward changes at the word, such as the removal of its suffix (MANNING; RAGHAVAN; SCHÜTZE, 2010). The objective of applying stemmization after a lemmatization was to ensure that combinations of similar meaning words transformed into a unique stem or base form. As such, each word of a publication t was transformed, going from $t = [w_1, w_2, \dots, w_s]$ to $t = [w'_1, w'_2, \dots, w'_s]$, where each $w'_i \subseteq w_i \quad \forall\, w_i \in t$. At the end of this phase each publication has its multiset of preprocessed tokens, which is used to construct the temporal graphs in the next phase.

3.4 PHASE 3: TEMPORAL GRAPH CONSTRUCTION

The third phase in this approach is related to Tasks 3, 4, and 5 from Figure 3.1. In this stage, we construct temporal graphs based on the T_{P_i} multisets obtained in the previous phase. The goal of these graphs is to model temporal variations in the words employed by users of Social Networks between two time windows since we want to verify changes in behavior from these users between such windows. To achieve this goal, we create a new set W with all preprocessed and unique words utilized by users on all publications from two consecutive time windows. For example, by comparing a pair of windows, there are $W' = \{w'_1, w'_2, \cdots, w'_{s'}\}$ and $W'' = \{w''_1, w''_2, \cdots, w''_{s''}\}$, each containing all words used in its window. Then, a set $W = W' \cup W''$ is created and, as stated in the Set Theory, must be composed of unique elements, i.e., $\forall w \in W$, $\exists! w \in W' \cup W''$. The cardinality $|W| = s$ of words in the set was applied to create an adjacency matrix $\mathbf{A}_{s \times s}$ (Task 3), employed when creating the graphs of this phase.

The first graph, Task 4 of Figure 3.1, represents the first time window and is created by counting the co-occurrence, i.e., appearance in the same publication, of every pair of words in its window. Thus, an element $a_{v,u} > 0$ of $\mathbf{A}$ represents an edge connecting two vertices v and u, i.e., $a_{v,u}$ counts all co-occurrences of the words v and u in a single window. At the end of this task, we have $G = \langle V, E \rangle$, where V is a set of words (vertices) used at least once during the two time windows, and E is a set of edges connecting those vertices but only considering the publications of the first one, as seen in Figure 3.3a. Next, $G' = \langle V, E' \rangle$ is created which represents the graph for the second window, Task 5 of Figure 3.1, where V contains the same words from G and E' connects these words considering only publications from the second time window, as seen in Figure 3.3b.



(a) Graph G of the first time window (P_1).

(b) Graph G' of the second time window (P_2).

Figure 3.3: Temporal graph from Phase 3, Tasks 4 (a) and 5 (b), of our approach.

As one may notice, for every time window, a graph is created respecting the following definitions: (i) there is no self-loop, i.e., a word connected to itself ($a_{v,v} = 0$); (ii) it is undirected; and (iii) it is weighted. An alternative method to represent the temporal graph built in this phase is through an adjacency matrix, as exemplified by Table 3.1 depiction of Figure 3.3.

Vertex	deuxxx		glor		jairbolsonar		daciolocab	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
deuxxx	0	0	11	14	10	9	20	0
glor	11	14	0	0	8	5	12	0
jairbolsonar	10	9	8	5	0	0	8	0
daciolocab	20	0	12	0	8	0	0	0

Table 3.1: Adjacency matrix for the temporal graph in Figure 3.3.

This phase is the most relevant contribution of our work since it models Social Network data into temporal graphs, which in turn depicts variations in their user's behavior between time windows. The temporal graphs generated at this phase were further explored, as seen in the next section, with the usage of concepts from the Time Series field.

3.5 PHASE 4: TIME SERIES EXTRACTION AND ANALYSIS

3.5.1 Extraction with Temporal Novelty Quantification

In this phase, at Task 6, we could use other measures, such as the topological overlap, to extract information from the previous $\widehat{G} = \{G_{P_1,P_2}, G_{P_2,P_3}, \ldots, G_{P_{h-1},P_h}\}$ temporal graphs and create a time series from them. Using topological overlap equation (Equation 2.3), on the graphs in Figures 3.4 and 3.5, we extract the data[1] seen in Table 3.2.

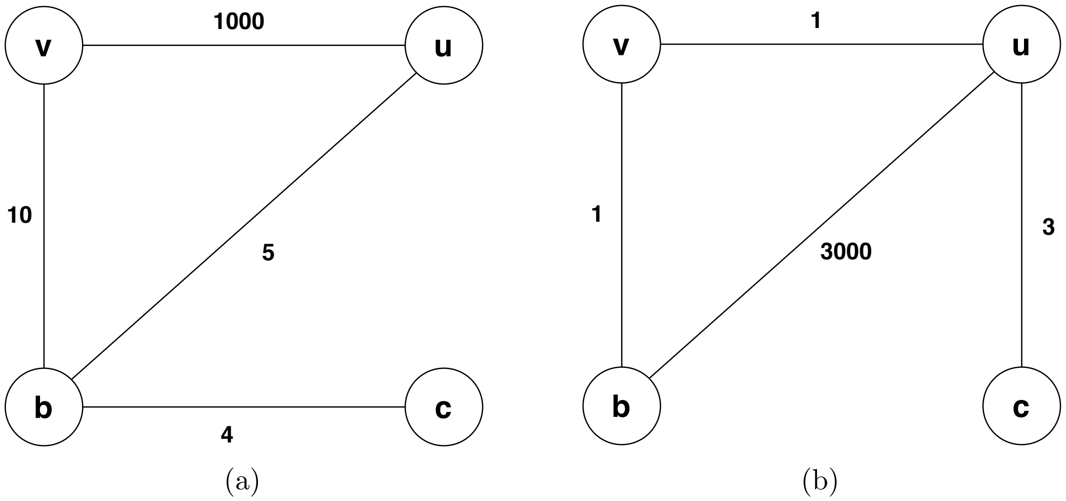


Figure 3.4: Temporal graph on instants P_1 (a) and P_2 (b).

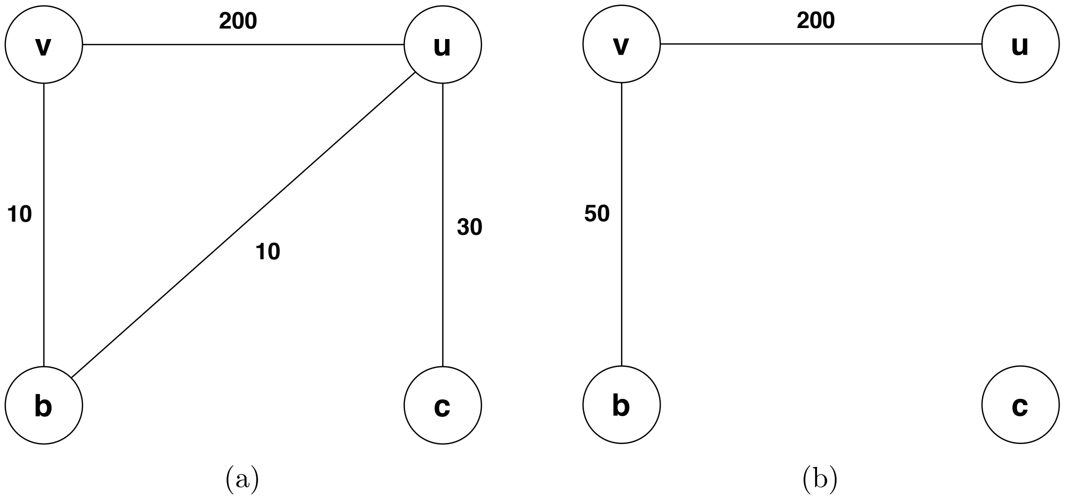


Figure 3.5: Another temporal graph on instants P_3 (a) and P_4 (b).

[1] In this work, we are using "." as a decimal separator and "," as a thousand separator.

Vertex	Topological Overlap		
	$P_1 \to P_2$	$P_2 \to P_3$	$P_3 \to P_4$
v	1	1	1
u	1	1	≈ 0.57
c	0	0	0
b	≈ 0.82	1	≈ 0.7

Table 3.2: Topological Overlap for each vertex of the graphs from Figures 3.4 and 3.5

This information achieves the measure goal of providing a probability in which a vertex edge occurs on the temporal graph and can be used to generate a time series. However, since topological overlap does not capture the weights of the edges, it is not clear when there is variation in the word relationship over time, which led us to create a new measure that depicts these changes.

First, we define a distance, in Equation 3.1, between nodes of a temporal graph to measure the relationship variation among these words. This equation uses a square root to limit the range of outputs and works as an inverse distance relation between words, i.e., the larger is their relationship in publications on both instants of the temporal graph, the shorter is their distance and vice versa.

$$\begin{aligned} d(u,v) &= \frac{|a_{u,v}(G) - a_{u,v}(G')|}{\sqrt{|a_{u,v}(G) - a_{u,v}(G')|}} \\ &= \sqrt{|a_{u,v}(G) - a_{u,v}(G')|} \end{aligned} \tag{3.1}$$

The sample temporal graphs from Figures 3.4 and 3.5 are used as examples, each depicting a pair of consecutive time windows. In Figure 3.4, when considering the transition between instants P_1 and P_2, edge (u, c) gets added, (c, b) is removed, and the last three change their weights. Meanwhile, in Figure 3.5, edge (v, u) stays the same, and (v, b) receives the weights from (u, c) and (u, b), which are removed. Table 3.3 presents the results of applying Equation 3.1 into the three transitions between time instants represented in these figures.

This equation provides three different interpretations regarding two words and the underlying topics which connect them: (i) if a pair of words have a strong connection during a time window, but it decreases significantly in the next window, then the topics have also diminished in importance; (ii) if there is an inverse behavior, then more publications are using these words together, which indicates an increase in the relevancy of the topics; finally, (iii) if the co-occurrence of some words remains similar between time intervals, then the topics significance is conserved. Using the previous graphs, we can see examples of such interpretations, with edges (v, u) and (u, b) from Figure 3.4 and edge (v, u) from Figure 3.5, respectively depicting each of the three possibilities. As one may notice, the first two interpretations are useful to identify when new behavior occurs, meaning that they can act as novelty detection.

Edge	Distance		
	$P_1 \to P_2$	$P_2 \to P_3$	$P_3 \to P_4$
(v,u)	$\sqrt{\lvert 1000-1 \rvert} \approx 31.6$	$\sqrt{\lvert 1-200 \rvert} \approx 14.1$	$\sqrt{\lvert 200-200 \rvert} = 0$
(v,b)	$\sqrt{\lvert 10-1 \rvert} = 3$	$\sqrt{\lvert 1-10 \rvert} = 3$	$\sqrt{\lvert 10-50 \rvert} \approx 6.32$
(v,c)	0	0	0
(u,b)	$\sqrt{\lvert 5-3000 \rvert} \approx 54.72$	$\sqrt{\lvert 3000-10 \rvert} \approx 54.68$	$\sqrt{\lvert 10-0 \rvert} \approx 3.16$
(u,c)	$\sqrt{\lvert 0-3 \rvert} \approx 1.73$	$\sqrt{\lvert 3-30 \rvert} \approx 5.2$	$\sqrt{\lvert 30-0 \rvert} \approx 5.47$
(c,b)	$\sqrt{\lvert 4-0 \rvert} = 2$	0	0

Table 3.3: Distances for each edge of the graphs from Figures 3.4 and 3.5.

With this distance, we can detect novelty in the relationship of a pair of words on the temporal graph. However, our goal at this phase is to analyze the behavior of users from Social Networks, which means the analysis needs to take into account all pairs of words used by them. To reach this objective, we define a new measure, in Equation 3.2, called Temporal Novelty Quantification (TNQ) ($\aleph$), which calculates the aggregate distance for all pairs of words, essentially quantifying the relationship variation between words employed in publications from two consecutive time windows. As an example, Table 3.4 portrays the aggregate values for the Temporal Novelty Quantification when considering the distances in Table 3.3.

$$\aleph(G, G') = \sum_{u,v \in G \bigcup G',\, u \neq v} d(u,v) \tag{3.2}$$

Temporal Graph	TNQ
$P_1 \to P_2$	93.05
$P_2 \to P_3$	76.98
$P_3 \to P_4$	14.95

Table 3.4: TNQ ($\aleph$) for each temporal graph in Figures 3.4 and 3.5.

The presented measure and its distance are relevant parts of this work and constitute one of our contributions. Another important aspect from them is their focus on transitions between periods $\{P_1 \to P_2, P_2 \to P_3, \ldots, P_{h-1} \to P_h\}$ instead of concentrating on a single instant like $\{P_1, P_2, \ldots, P_h\}$. Still, both contain limitations, with the measure obfuscating individual contributions from pairs of words and the distance ignoring the trend of the variation, i.e., if a relationship between words decreases rapidly, the distance increases. Regardless of their faults, after applying them on every temporal graph generated in the previous phase, we obtained a sequence of values $Y = (y_{1,2}, y_{2,3}, \ldots, y_{h-1,h})$ with each representing a transition between two periods. This sequence of values is then able to be modeled as a time series, as seen in Figure 3.6, which uses data from Table 3.4.

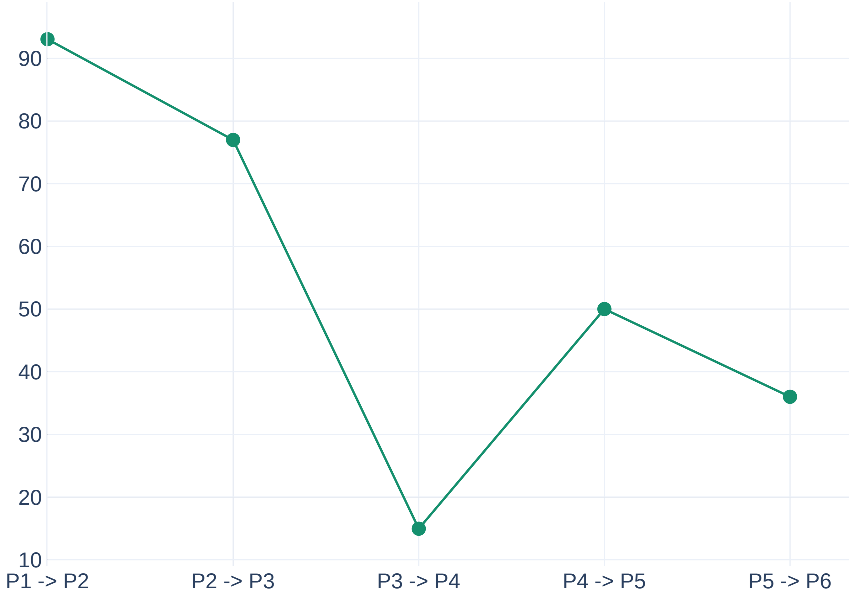


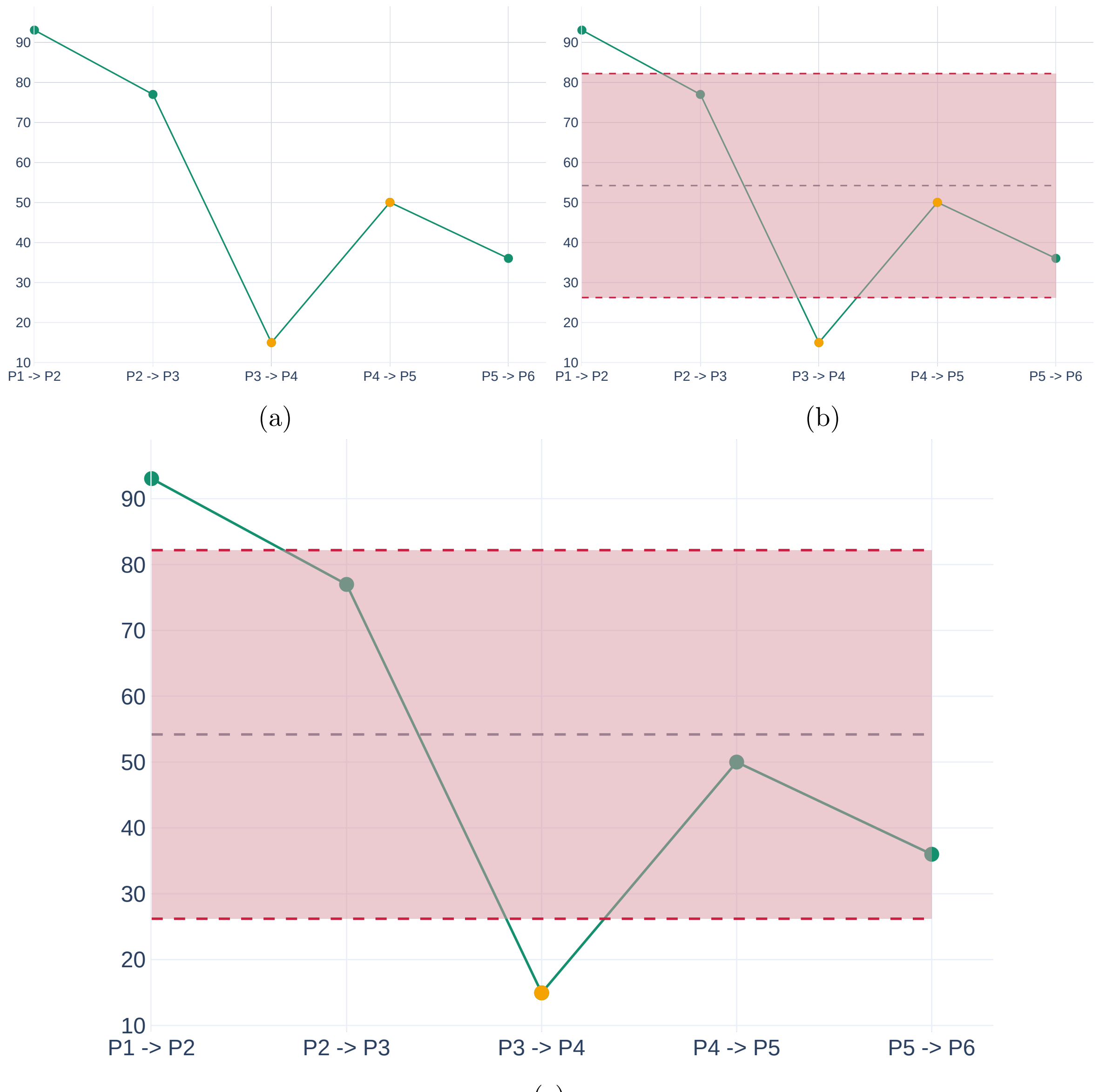
Figure 3.6: Time series generated from Temporal Novelty Quantification data on Table 3.4.

3.5.2 Analysis with Concept Drift

After the extraction of this time series, which models Social Network user's behaviors through variation in the words used by them over time, it is possible to apply different tools from the Time Series field to obtain useful information from it. Since the primary goal of this work is to detect changes in users' behaviors, we decided to use a Concept Drift (SCHLIMMER; GRANGER JR., 1986; WIDMER; KUBAT, 1996; GAMA et al., 2014) technique, in Task 7 of our approach, identifying changing points in the series that indicate these variances, instead of looking at every observation in the series.

Through an examination of traditional Concept Drift techniques presented by Gama et al. (2014), we noticed that some methods, such as STAGGER (SCHLIMMER; GRANGER JR., 1986) or FLORA (WIDMER; KUBAT, 1996), assume two characteristics from the data they analyze. First, they require a large number of observations with defined time windows in which to detect possible changes. Secondly, they assume each instant in the series has a classification label or will be classified since their application is within a supervised Machine Learning setting. This last assumption is unfeasible in our current scenario, preventing our usage of these techniques, which led to the development of a new straightforward Concept Drift algorithm to support us in achieving our final research goal.

In our method, we used the algorithm defined on (KIM; OH, 2009; KIM; OH, 2018) for local extrema detection, which are points of maximum or minimum values present in time series (FLANIGAN et al., 1998). This algorithm uses the difference, and its sign, between consecutive instants in the time series and then verify if its tendency has changed, which characterizes an extremum. Figure 3.7a displays all extrema detected, based on the time series from Figure 3.6, as yellow dots with their values being 14.95 and 50. Still, since user behavior is very dynamic, it is not desirable to consider each small change in the series behavior, driving the creation of a tolerance interval in which changes would be regarded as natural. Defined as $[\mu - \sigma, \mu + \sigma]$, based on the series mean (μ) and standard deviation (σ), this interval is show in Figure 3.7b as the pink area. Finally, our last step is then to select the extrema outside the tolerance interval, i.e., $\forall y_i \in Y$ such that $y_i \not\subset [\mu - \sigma, \mu + \sigma]$, which are the remaining yellow dots in Figure 3.7c.



(a)

(b)

(c)

Figure 3.7: Steps to identify Concept Drift points: Extrema identification (a), Interval definition (b) and Extrema selection (c).

With the finding of these spots, our approach identifies moments of concept drift in the behavior of users from Social Networks, achieving the goal of this work. Further investigations on these moments, e.g., what or who caused them, are outside the scope of our designed approach, yet, the next chapter presents a case study that performs further examinations regarding the collected concept drifts.

3.6 FINAL REMARKS

This chapter detailed all steps performed by our approach. In brief, it starts transforming a set of publications (Tasks 1 and 2 in Figure 3.1) from a Social Network into a Time Varying Graph built on top of time windows (Tasks 3 to 5). Then, the Temporal Novelty Quantification is used to measure the relationship between pairs of graphs. As a consequence of applying the proposed quantification on such graphs, a time series is produced (Task 6), thus allowing a better understanding of people's reactions on specific topics. For example, it is possible to use time series methods (Task 7) to detect trends, spikes, bottoms, or, using our interpretation, changing points. The next chapter presents a case study, which was carried out to assess our approach, that analyzes and correlates the points found with real-world events.

Chapter

4

CASE STUDY: BRAZILIAN POLITICS IN 2018 ELECTION

4.1 INITIAL REMARKS

This chapter presents experiments performed to evaluate our approach, based on data collected from Twitter in 2018. It is important to recall our methodology steps to model such data: i) Social Network data collection; ii) Preprocessing; iii) Temporal graphs construction; and iv) Time series extraction and analysis. In the next section, we present a brief overview on the setup used to execute these steps. Our datasets, source code, and a step-by-step on how to reproduce our results are accessible in a public repository [1], with all of them open to improvements and free to use.

4.1.1 Experimental Setup

This experiment was executed on a cluster called "Euler" located at the University of São Paulo (USP), which has the following elements:

(i) CPU: 98.4 TFLOPS;

(ii) GPU: 28 TFLOPS; and

(iii) XeonPhi: 1,011 GFLOPS.

In summary, its maximum theoretical performance is close to 127 TFLOPS. For sequential processing, this cluster provides 4 nodes with 2 Intel Xeon E5-2667v4 3.2 GHz processors with 8 cores and dedicated memory (512 GB DDR3 1866MHz). For parallel processing, there are 6,400 Hyper-threading processors and 18 TB of shared memory. Before presenting the experiment results, we show practical examples to illustrate every step of our approach, as previously examined in Chapter 3 and Figure 3.1.

[1] Available at: ⟨https://gitlab.com/m1thr4nd1r/social-network-analysis⟩

4.2 DATA COLLECTION PHASE (TASK 1)

As aforementioned, we used texts published on Twitter in October 2018 as a proof of concept for our approach. This period, marked by the Brazilian presidential election, presented an expressive number of political publications, thus definitely establishing Social Networks as a new and influential political advertising and activism method in Brazil.

The political polarization in the country motivated us to monitor the most influential politicians in the 2018 election race: the current and a former president, Jair Bolsonaro (@jairbolsonaro) and Lula (@LulaOficial), respectively. Although the Superior Electoral Court has barred Lula from the presidential race on August 31st, 2018, his name remained in people's minds during almost the whole election. In this sense, for each politician, we filtered any tweet in Portuguese that directly mentions its username (e.g., "@LulaOficial") or that has a hashtag equal to its username (e.g., "#jairbolsonaro") and used the TSViz (RIOS et al., 2017a; RIOS et al., 2017b; MELLO et al., 2018) platform to collect these tweets. Figure 4.1 summarizes the publication volume and calls attention to the daily number of tweets (y-axis). Tweets related to former president Lula show a spike with more than 30k tweets on the first election round eve (6th). Concerning the current president Bolsonaro, the astounding volume of tweets published during the day of second-round voting (28th) has reached almost 450k tweets.

4.3 PREPROCESSING PHASE (TASK 2)

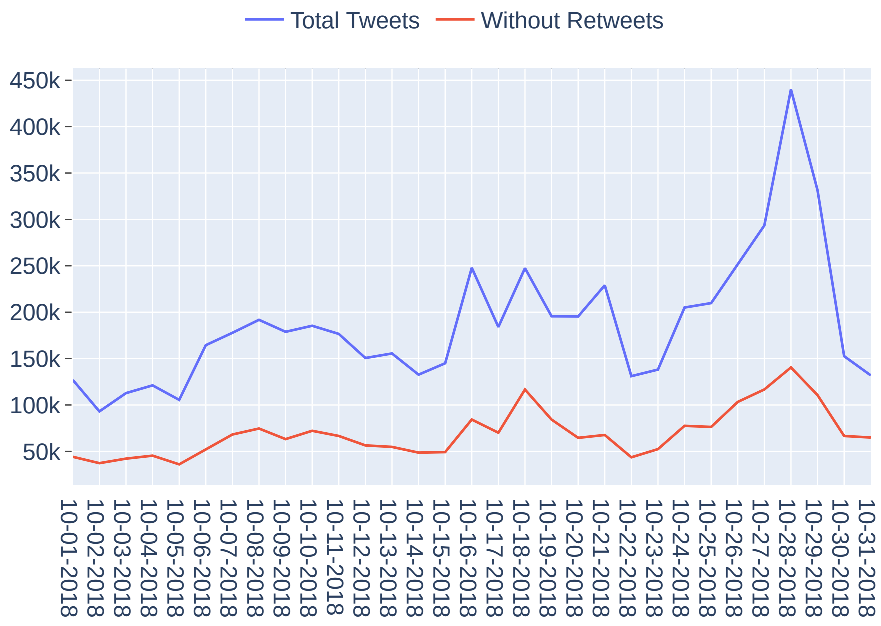
After collecting the tweets, the first steps done in the preprocessing were to remove any tweet starting with "RT" and to organize texts by a period $P \subseteq \Delta$. On Twitter, the term "RT" stands for retweet, which implies the publishing of the same content previously written by someone else.

The original volume of tweets for each observed politician and the corpus, which corresponds to the whole set of texts analyzed, is seen in Table 4.1. This table also shows the final volume of tweets after removing RTs, revealing a significant reduction for both politicians, which is seen in the distance between the blue and red lines in Figure 4.1.

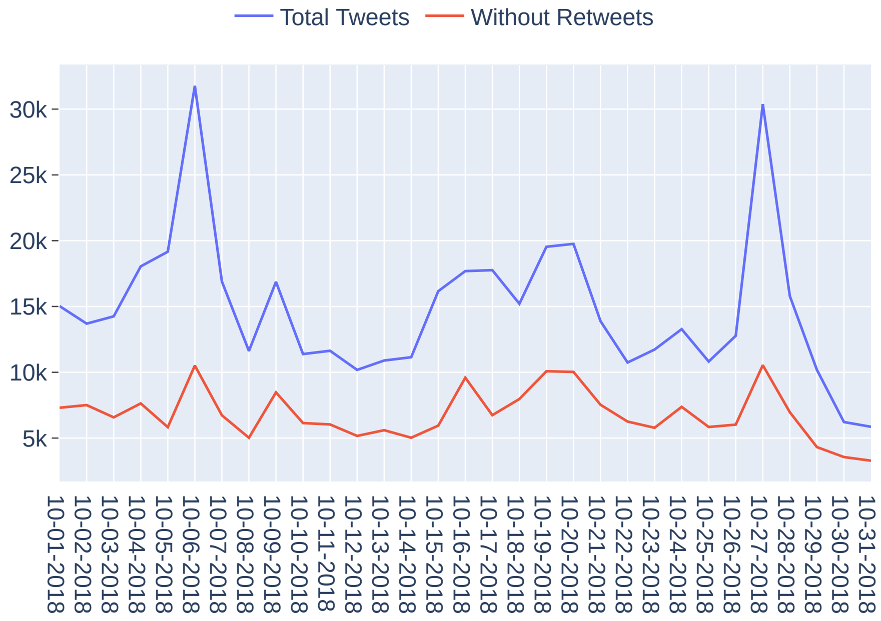
Topic	Amount of Tweets		
	Total	Without Retweets	Decrease (%)
Bolsonaro	5,802,028	2,150,306	3,651,722 (≈ 63%)
Lula	460,379	211,386	248,993 (≈ 54%)
Corpus	6,262,402	2,361,692	3,900,715 (≈ 62%)

Table 4.1: Amount of tweets before and after preprocessing of the corpus.

At this point, the remaining tweets still need to be preprocessed before being used in the construction of temporal graphs. To better exemplify the next steps, we selected the following tweets as examples: *"@allnicksused Eles amam o @jairbolsonaro kkkkkk #DebateNaRecord"* and *"@jairbolsonaro Olha só que lindo lindo 👏👏👏👏👏👏👏👏👏👏👏👏👏*



(a) Bolsonaro

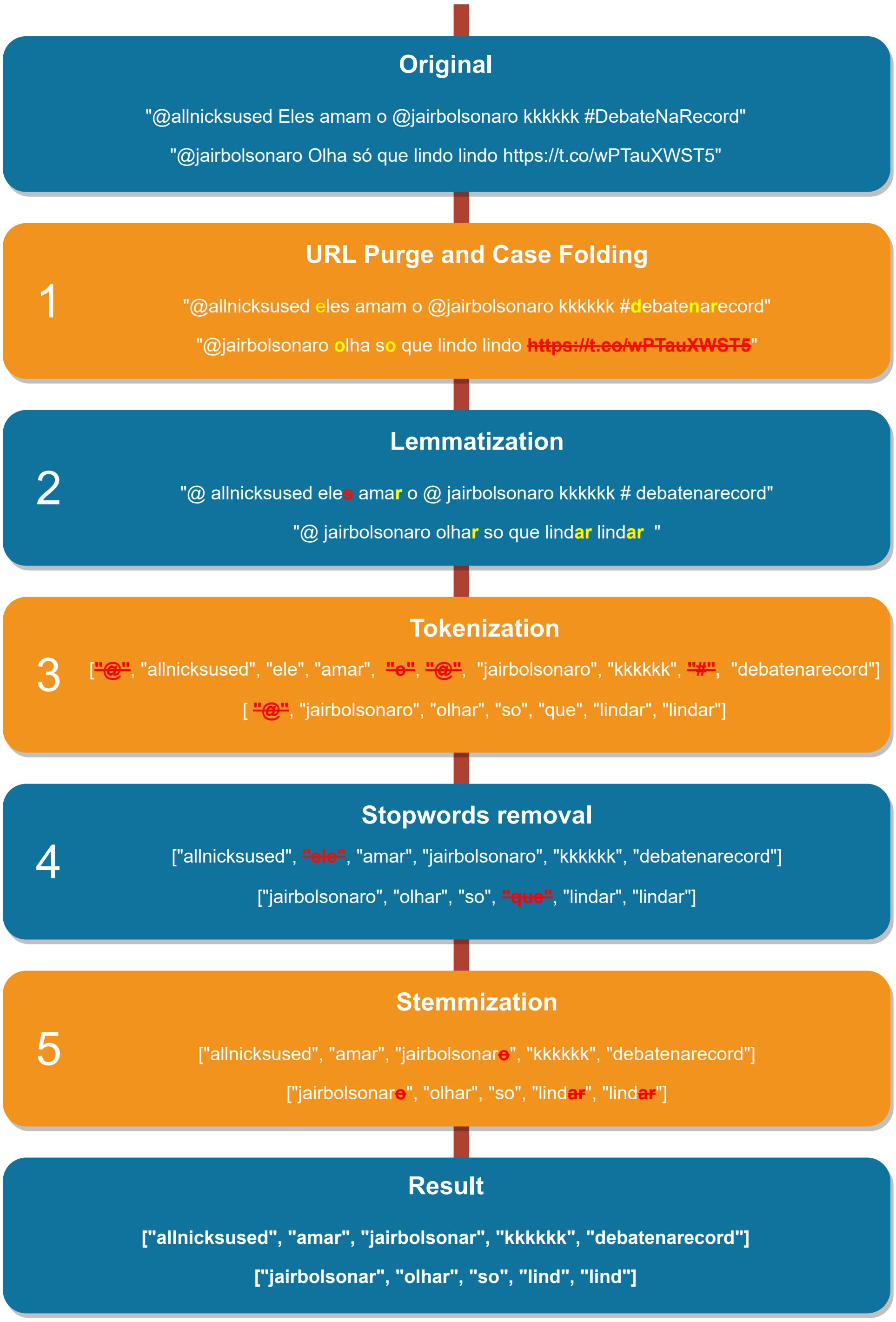


(b) Lula

Figure 4.1: Volume of published tweets related to Bolsonaro (a) and Lula (b).

⟨https://t.co/wPTauXWST5⟩"[2]. Next, we discuss every preprocessing performed on them, as summarized in Figure 4.2.

[2] The tweets used are in Portuguese due to the focus of this research. Still, the reader can follow the same preprocessing steps regardless of the language in its texts.



Original

"@allnicksused Eles amam o @jairbolsonaro kkkkkk #DebateNaRecord"

"@jairbolsonaro Olha só que lindo lindo https://t.co/wPTauXWST5"

1 URL Purge and Case Folding

"@allnicksused eles amam o @jairbolsonaro kkkkkk #debatenarecord"

"@jairbolsonaro olha so que lindo lindo ~~https://t.co/wPTauXWST5~~"

2 Lemmatization

"@ allnicksused eles amar o @ jairbolsonaro kkkkkk # debatenarecord"

"@ jairbolsonaro olhar so que lindar lindar "

3 Tokenization

[~~"@"~~, "allnicksused", "ele", "amar", ~~"o"~~, ~~"@"~~, "jairbolsonaro", "kkkkkk", ~~"#"~~, "debatenarecord"]

[~~"@"~~, "jairbolsonaro", "olhar", "so", "que", "lindar", "lindar"]

4 Stopwords removal

["allnicksused", ~~"ele"~~, "amar", "jairbolsonaro", "kkkkkk", "debatenarecord"]

["jairbolsonaro", "olhar", "so", ~~"que"~~, "lindar", "lindar"]

5 Stemmization

["allnicksused", "amar", "jairbolsonar~~o~~", "kkkkkk", "debatenarecord"]

["jairbolsonar~~o~~", "olhar", "so", "lind~~ar~~", "lind~~ar~~"]

Result

["allnicksused", "amar", "jairbolsonar", "kkkkkk", "debatenarecord"]

["jairbolsonar", "olhar", "so", "lind", "lind"]

Figure 4.2: Example of every preprocessing step applied document-wise.

4.3.1 URL Purge and Case Folding

This step removes any URL or external link from a tweet t, optimizing data for the next preprocessing steps. Usually, these are references to news, images, videos, or other multimedia content that is not available on Twitter. To achieve this goal, we use a regular expression, or regex, to remove these references from a tweet since they do not bring any information by themselves. The expression used [3] is not presented here, mainly due to its high complexity, which would lead us outside the scope of this work to explain it.

At this stage, a case folding is performed to lowercase, thus avoiding case-sensitive issues, such as "god" and "God" being considered different words. Moreover, we also removed all symbols, emojis, and word accents, with Table 4.2 showing the daily average for URLs and word accents removed for each politician, the complete transformation performed in this step is illustrated below and in Step 1 (Figure 4.2).

- "@allnicksused Eles amam o @jairbolsonaro kkkkkk #debatenarecord" becomes "@allnicksused eles amam o @jairbolsonaro kkkkkk #debatenarecord";

- "@jairbolsonaro Olha só que lindo lindo https://t.co/wPTauXWST5" turns into "@jairbolsonaro olha so que lindo lindo ".

Politician	Daily Average	
	Removed URLs	Removed Word Accents
Bolsonaro	$\approx 24,258$	$\approx 152,096$
Lula	$\approx 2,570$	$\approx 13,384$

Table 4.2: Daily average of removed URLs and word accents for Bolsonaro and Lula.

4.3.2 Lemmatization

In this step, we apply a lemmatization (MANNING; RAGHAVAN; SCHÜTZE, 2010) tool to reduce the words to their grammatical root. This tool is called "lematizador" and was created by the Interinstitutional Nucleus of Computational Linguistics (NILC) lab (NUNES et al., 1996) from USP, with Step 2 (Figure 4.2) illustrating its application to the previous tweets, also replicated next:

- "@allnicksused eles amam o @jairbolsonaro kkkkkk #debatenarecord" changes to "@ allnicksused ele amar o @ jairbolsonaro kkkkkk # debatenarecord";

- "@jairbolsonaro olha so que lindo lindo " transforms into "@ jairbolsonaro olhar so que lindar lindar ".

[3] Available at: ⟨https://gitlab.com/m1thr4nd1r/social-network-analysis/-/blob/master/src/preprocessing.py#L12-14⟩

4.3.3 Tokenization

This stage also uses a regular expression to tokenize (WEBSTER; KIT, 1992) the data, thus separating the words contained in them, which is the following:

$$\backslash \mathrm{bw}\{2,\}\backslash \mathrm{b}$$

In this regular expression, "\b" matches with a word boundary, i.e., the position when a word starts or ends, which can be a non-alphanumeric character, such as a hashtag (#). When using an inner expression between word boundaries, we are effectively delimiting specific sequences within them. In this case, the regex "w{2,}" looks for two or more alphanumeric characters, such as letters or numbers. The output of this regular expression on the previous tweets is presented next and illustrated in Step 3 (Figure 4.2):

- "@ allnicksused ele amar o @ jairbolsonaro kkkkkk # debatenarecord" is split in "allnicksused", "ele", "amar", "jairbolsonaro", "kkkkkk", "debatenarecord";
- "@ jairbolsonaro olhar so que lindar lindar " is tokenized as "jairbolsonaro", "olhar", "so", "que", "lindar", "lindar".

4.3.4 Stopwords Removal

The next step was to remove stopwords (LUHN, 1960), which are terms that do not bring any new knowledge to an analysis. In our context, the stopwords removed were based on the Portuguese ones found at the Natural Language Toolkit (NLTK) (LOPER; BIRD, 2002) package. The subset of remaining words, after the removal of stopwords from the previous tweets, can be seen below and in Step 4 (Figure 4.2):

- "allnicksused", "ele", "amar", "jairbolsonaro", "kkkkkk", "debatenarecord" is reduced to: "allnicksused", "amar", "jairbolsonaro", "kkkkkk", "debatenarecord";
- "jairbolsonaro", "olhar", "so", "que", "lindar", "lindar" results in: "jairbolsonaro", "olhar", "so", "lindar", "lindar".

4.3.5 Stemmization

The final preprocessing step was the execution of a stemmization (LOVINS, 1968) algorithm, as seen in Step 5 of Figure 4.2. We used the Snowball Portuguese Stemmer (PORTER, 2001), and the results of its application to the previous tweets are:

- "allnicksused", "amar", "jairbolsonaro", "kkkkkk", "debatenarecord" stemmizes to: "allnicksused", "amar", "jairbolsonar", "kkkkkk", "debatenarecord";
- "jairbolsonaro", "olhar", "so", "lindar", "lindar" changes to: "jairbolsonar", "olhar", "so", "lind", "lind".

After all these modifications, the last box in Figure 4.2 shows the final lists of terms for both tweets. Meanwhile, in Table 4.3, we see the reduction in the average of unique words per day for both politicians after each preprocessing step applied to the tweets, thus decreasing the amount of information used and the resources needed to create the temporal graphs in the next phase

Daily Average of Unique Words	Politician	
	Bolsonaro	Lula
Raw Tweets	$\approx 59,020$	$\approx 12,899$
After Lemmatization	$\approx 47,592$	$\approx 10,595$
After Tokenization	$\approx 47,005$	$\approx 10,494$
After Stopwords Removal	$\approx 46,930$	$\approx 10,433$
After Stemmization	$\approx 37,174$	$\approx 8,550$

Table 4.3: Daily average of unique words in each preprocessing step for each politician.

4.4 TEMPORAL GRAPH CONSTRUCTION (TASKS 3, 4 AND 5)

After preprocessing all tweets related to Bolsonaro and Lula published between $\Delta = \{01\text{ October 2018}, 31\text{ October 2018}\}$ and considering the political moment of the country, we defined the time window as one day ($P = 1$), that means the temporal relation in this experiment was obtained by comparing pairs of sequential days.

The previous chapter explained the build process of the temporal graphs, but in summary, for each graph, words from its pair of consecutive days are used as nodes while their edges represent the relationships of these words in each day. After building a temporal graph for each duo of sequential days, we obtain a list of temporal graphs $\widehat{G} = \{G_{1,2}, G_{2,3}, \ldots, G_{h-1,h}\}$ spanning the whole observed month.

Information of the temporal graphs for Bolsonaro and Lula, between the 1st and the 2nd, is displayed in Table 4.4. It shows the expressive number of vertices and edges for a single temporal graph, which is not the one with the most nodes or edges.

Politician	Temporal Graph ($G_{1,2}$)			
	First Instant Graph (G_1)		Second Instant Graph (G_2)	
	Vertices	Edges	Vertices	Edges
Bolsonaro	$34,970$	$598,377$	$34,970$	$706,751$
Lula	$13,639$	$212,988$	$13,639$	$207,486$

Table 4.4: Amount of vertices and edges for a temporal graph between days 1 (G_1) and 2 (G_2) for Lula and Bolsonaro.

The size of these graphs also demonstrates the difficulty in visually representing them, even if we were to extract only a limited subset of nodes and edges, which would require the definition of criteria to select them. Nevertheless, a visualization of a sample temporal graph is available in Figure 3.3. Thus, with the graphs constructed, we can proceed with the extraction of the time series and its analysis, as seen next.

4.5 TIME SERIES EXTRACTION AND ANALYSIS (TASKS 6 AND 7)

One way to extract information from temporal graphs is through the usage of measures, which can be specific to them (e.g., Topological Overlap) or derived from measures for static graphs (TANG et al., 2010a; KIM; ANDERSON, 2012). In this work, due to some limitations presented on such measures in the previous chapters, we used a new one, called Temporal Novelty Quantification (TNQ) ($\aleph$). When applied to the $\widehat{G}$ set of temporal graphs constructed previously, we extract a sequence of scalar values that organizes users' behavior variation, through words used by them throughout the whole time Δ observed.

We arranged this sequence as a time series. Although it is possible to analyze every instant of it, our final goal is to identify when behavior changes happen on users of Social Networks, which drove our usage of Concept Drift techniques on time series. Still, in general, these techniques have some requirements that prevent their application in this work, which directed us to create a new straightforward algorithm for its detection. It selects all local extrema, using the algorithm defined in (KIM; OH, 2018), that is outside a tolerance interval, based on the series mean and standard deviation, as concept drifts.

Figures 4.3 and 4.4 depict the time series and concept drifts found for Bolsonaro and Lula, respectively, with Table 4.5 highlighting the selected points.

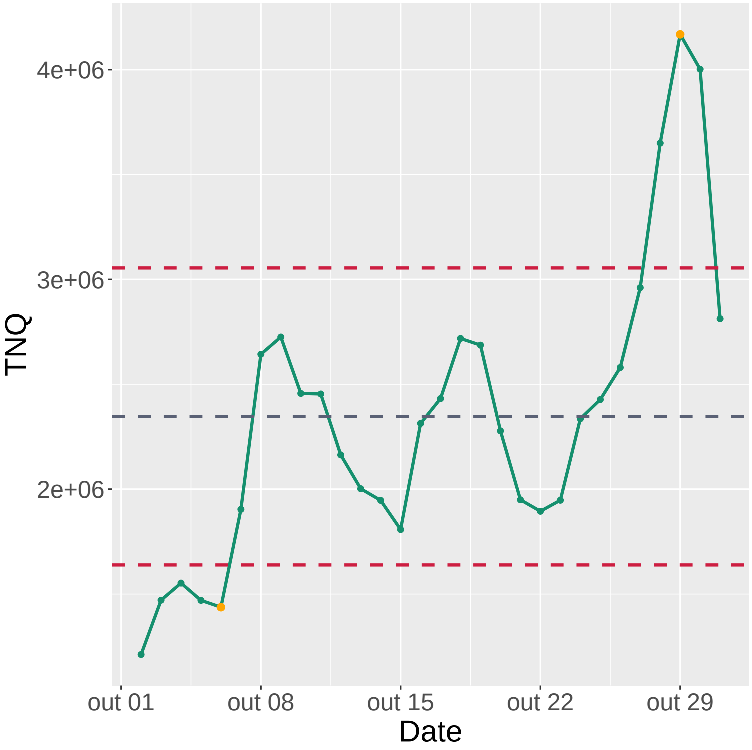


Figure 4.3: Time series produced after analyzing tweets with reference to Bolsonaro.

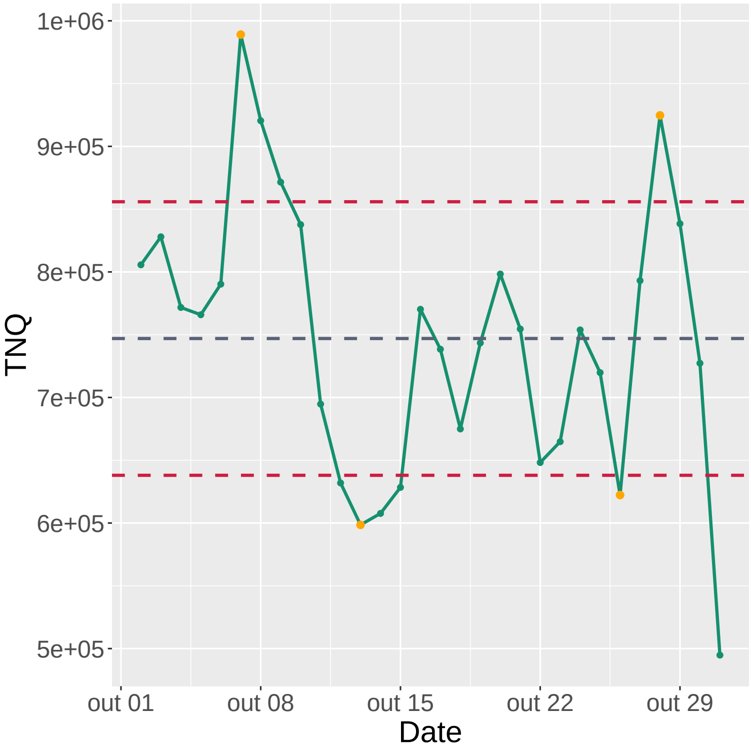


Figure 4.4: Time series produced after analyzing tweets with reference to Lula.

Politician	Concept Drifts (days)
Bolsonaro	6, 29
Lula	7, 13, 26, 28

Table 4.5: Concept Drifts found for Bolsonaro and Lula.

The next section investigates these points for each of them, looking at what events and headlines made the news, especially at the political level and involving the politician, that could suggest why users' behavior has changed.

4.6 RESULTS

The volume of tweets published during October 2018 in Brazil (Figure 4.1) emphasizes the use of Social Networks as new mechanisms of political activism. Recently, it has been common to notice police investigations connecting such platforms to bots designed to spread fake news and influence political elections. The volume of collected data and its reduction to 1/3 of the original size after removing retweets, shown in Table 4.1, indicate Twitter has more usage in promoting someone rather than to discuss its points of view.

The time series produced after analyzing the temporal graphs based on tweets published with some reference to Bolsonaro, without retweets, is displayed in Figure 4.3, with two dates highlighted after the execution of our Concept Drift method: 6 and 29th of October

2018. We proceed to examine what was happening in the Brazilian political scenario related to him that could cause changes in the behavior of users from Twitter.

Analyzing the region, one day after and before the point, that is 5, 6, and 7, we noticed how they are close and even overlaps with the voting day of the initial round of elections. Voters can decide the election at this instant, which makes it the most important moment so far, with users expected to change their behaviors and intensify their activism to support their candidates as much as possible. Regardless, favoritism for Bolsonaro, especially after an attempt against his life, led to expressive results in this first round for him, which called the attention of traditional media [4,5,6].

Meanwhile, the second region, i.e., days 28, 29, and 30, corresponds to the final election day and its following days. At this point, Bolsonaro is announcing plans for his government [7], such as inviting Sergio Moro to be the Minister of Justice and Public Security [8]. Sergio Moro gained fame with Operation Car Wash, which curiously even investigated the former president Lula. With the winner proclaimed and starting to announce his projects for the country, users could reduce their political activism and react to such plans, reducing the TNQ values.

A time series for Lula, generated in the same form as Bolsonaro's but with a different set of tweets, is shown in Figure 4.4, with yellow dots marking the four emphasized points after the application of our Concept Drift technique. Then, we inspected the regions centered on these days while looking for events related to Lula that could cause changes in users' behaviors.

Similar to the previous series, the initial region (i.e., days 6, 7, and 8) overlaps with the voting day for the first election round, with users more prone to talk about political matters these days. The voting results qualified Lula's substitute, Fernando Haddad, to the next voting round and established him as the remaining opposition to Bolsonaro. On the next day, Haddad goes to discuss his next steps towards the second round of elections with Lula [9], who was in jail, in an unusual event that reverberated with Twitter users.

The second region, centered in the concept drift at the bottom, encompasses days 12, 13, and 14th, starting right when both candidates' parties released new campaign advertisements using Lula's image. While materials for Haddad show Lula complimenting him, commercials for Bolsonaro keep associating Haddad with Lula and mentioning how they will bring communism to Brazil [10]. Meanwhile, on 14th, Bolsonaro once again indicates that Lula is commanding Haddad and that he refuses to take part in a debate if

[4] See:⟨https://oglobo.globo.com/brasil/tudo-sobre-candidato-presidencia-jair-bolsonaro-psl-23123698⟩

[5] See:⟨https://www.bbc.com/portuguese/brasil-45768006⟩

[6] See:⟨https://www.gazetadopovo.com.br/opiniao/artigos/o-fenomeno-bolsonaro-28wcdvyckyt4miedabe14zmxl/⟩

[7] See:⟨https://www.bbc.com/portuguese/brasil-46017462⟩

[8] See:⟨https://g1.globo.com/politica/noticia/2018/10/29/bolsonaro-diz-que-convidara-sergio-moro-para-ministro-da-justica-ou-o-indicara-para-o-stf.ghtml⟩

[9] See:⟨https://politica.estadao.com.br/noticias/geral,haddad-visita-lula-para-discutir-2-turno,70002538683⟩

[10] See:⟨https://oglobo.globo.com/brasil/videos-na-tv-bolsonaro-usa-lula-em-ataques-haddad-fala-em-casos-de-intolerancia-politica-23152195⟩

Lula participated [11], amplifying both the polarity of this election and TNQ values.

The final two regions, days 25, 26, and 27th for one and 27, 28, and 29th for the other, overlap between themselves, denoting the last stretch of the election race. On October 26th, another unexpected event happened when Roger Waters, an English singer, composer, and political activist, asked to visit Lula in jail [12], which again echoed with Twitter users and drove TNQ values up. The last concept drift point is related to the second election round on the 28th, which focused public opinion on Bolsonaro's win, leading to a drop in TNQ values for Lula but an increase for Bolsonaro, as seen in Figure 4.3.

4.7 FINAL REMARKS

This chapter covered in detail the experiment carried out to validate the approach designed, which aims to detect changes in the behavior of users from Social Networks. The base of such study case came from data related to two of the most popular candidates to the presidential office in the Brazilian election of 2018: Lula and Bolsonaro. Our approach obtained results closely associated with important events that happened during the election period, thus corroborating its ability to identify moments that could lead to some change in users' behavior. In the next chapter, we draw our conclusions and present some options for future works.

[11] See:⟨https://politica.estadao.com.br/noticias/eleicoes,eu-nao-iria-debater-com-lula-de-jeito-nenhum-afirma-bolsonaro,70002547290⟩

[12] See:⟨https://politica.estadao.com.br/noticias/eleicoes,roger-waters-pede-para-visitar-lula-na-prisao-em-curitiba,70002566404⟩

Chapter 5

CONCLUSION

5.1 DISCUSSIONS ABOUT OUR APPROACH

This dissertation presented our approach, designed to model textual information published on Social Networks and their relationships while tracking changes that occur on the texts or their relations. Since the information published on these platforms corresponds to their users' expressions, we could also analyze users' behavior and identify novelties on it through variations in the words used by them in an interval of time regarding specific subjects.

Initially, our approach applies traditional Text Mining techniques on the collected texts to filter terms written by users. Then, we construct Time Varying Graphs that keep the relation between tokens and their temporal dependencies. Next, we analyzed the temporal graphs with our new measure Temporal Novelty Quantification (TNQ), which models the variation between non-overlapping and consecutive time-windows, thus producing a time series. Subsequently, we devised a simple algorithm to detect changing points over the series. We want to highlight that our approach is flexible, which means that another research can freely and easily use it with a different preprocessing, temporal graph measure, Concept Drift algorithm, or correlation between time series behavior and news, if it so desires.

The results found in the experiments performed in the previous chapter shows that the concept drift points found by our approach, which indicates moments where the behavior of the users changed, highlights real-world events that might affect their behavior regarding a particular topic, e.g., politics. Thus, the designed approach was able to validate the hypothesis presented in this work, which was its primary aspiration while also achieving its specific objectives. Another significant achievement of this master work was the publication of a research paper, based on this approach and its obtained results, on the Brazilian Conference on Intelligent Systems (BRACIS) 2020 (SANTOS et al., 2020), a relevant national conference on Artificial Intelligence.

Nevertheless, we faced difficulties in this approach, mostly in Phases 2 and 3. The preprocessing phase takes more than a day to process one day, especially for Bolsonaro, where the busiest days can take two or three days to complete. This delay occurs mainly due to lemmatization, which also requires a large amount of system memory, e.g., 256 GB for one of the days with the highest volumes of tweets. Memory is also crucial in Phase 3 when we need to create a matrix of millions or even billions of entries (e.g., the square of unique words in Table 4.3), possibly requiring around 32 GB to process information for a single temporal graph. All these emphasize the greatest bottleneck of this project, which was, and still is, the availability of computational resources to run our approach in the volume of collected texts. Thus, we are thankful for the University of São Paulo (USP) and the São Paulo Research Foundation (FAPESP), with the project under grant number 2013/07375-0, for allowing us to use the cluster "Euler".

Moreover, during the experiments, we noticed an unusual number of texts, which were not retweets but verbatim copies of previous tweets. Although human beings are allowed to perform such behaviors, this is not common and relates more to the operation executed by bots developed to place users and events among the key (trending) topics discussed on Twitter. Figure 5.1 illustrates this situation, displaying the number of cloned tweets present on each day for Bolsonaro and Lula after the removal of retweets. Meanwhile, Table 5.1 shows for each of them, the average volume of cloned tweets and the highest amount of clones in a single day, which was day 27th for both.

Politician	Cloned Tweets	
	Average	Highest Volume
Bolsonaro	$\approx 3,860$	$10,271$
Lula	≈ 791	$1,801$

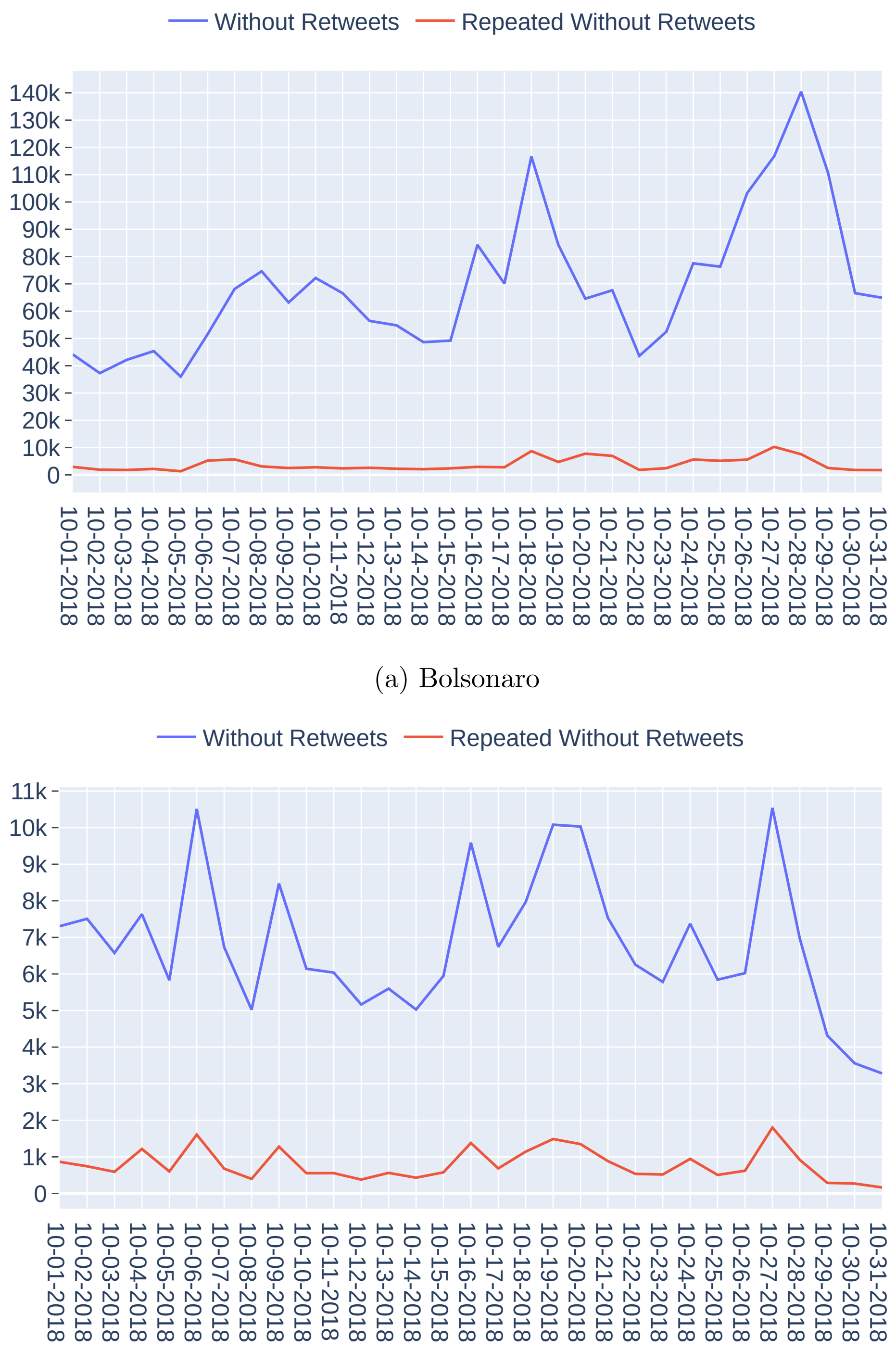
Table 5.1: Average and highest volume (27th) of duplicated tweets for both politicians.

All observations performed so far directed us to suggestions for future research that can be done based on this work, which we present next.

5.2 FUTURE WORK

During the development of this work, we have noticed improvements and other investigations that additional research could explore. The first one would be to investigate the impacts of each step in Phase 2 (e.g., what happens if the current preprocessing skips lemmatization) both in terms of time consumed and final results obtained with or without the investigated step.

Moreover, we remark that using entities instead of words as the most basic information could act as a filter to the terms used to construct the temporal graphs. Another form of filtering could come from the usage of other techniques from the Text Mining and Machine Learning fields, such as Term Frequency–Inverse Document Frequency (TF-IDF) and word clustering methods. These filters are appropriate because words used on Social Networks can radically change, even in a short time, leading to terms still impacting the analysis long after their relevance is gone.



(a) Bolsonaro

(b) Lula

Figure 5.1: Volume of cloned tweets, without retweets, from Bolsonaro (a) and Lula (b).

With this behavior in mind, the temporal graphs could have dynamic nodes, i.e., nodes that vary over the graph time interval, reflecting more faithfully the behavior of messages on Social Networks. Furthermore, the graphs generated contain amounts of edges and vertices that make their visualization unfeasible. Such visualization could help understand the relationships present on them, which makes this a relevant research possibility.

In our last phase, we understand that the investigation of TNQ values by users, separating their contribution on each tweet, could facilitate the identification of bots since users with high TNQ values are unlikely to be ordinary human users. Besides, a normalization factor for TNQ, e.g., dividing it by the mean number of tweets published between the two analyzed days, can be used to reduce the effect of the volume of publications whenever necessary. Other forms of concept drift detection, not using local extrema, can be investigated along with different analyses and ways to associate the concept drifts found with real-world events.

Another relevant future study would be the analysis of different politicians or subjects in conjunction, quantifying how they cite each other in their publications over time. This analysis puts the focus of the research into a subset of terms that occurs in a smaller group of users' publications, reducing the texts collected and the size of the temporal graphs. Finally, we suggest the investigation of more recent political events, such as the Brazilian election of 2020, or even other case studies in different areas such as environmental activism, minority underrepresentation, COVID-19, and the entertainment industry.

BIBLIOGRAPHY

ACQUISTI, A.; FONG, C. An experiment in hiring discrimination via online social networks. *Management Science*, v. 66, n. 3, p. 1005–1024, 2020. Available at: ⟨https://doi.org/10.1287/mnsc.2018.3269⟩.

AOUN, N. B.; MEJDOUB, M.; AMAR, C. B. Graph-based approach for human action recognition using spatio-temporal features. *J. Vis. Comun. Image Represent.*, Academic Press, Inc., Orlando, FL, USA, v. 25, n. 2, p. 329–338, fev. 2014. ISSN 1047-3203. Available at: ⟨http://dx.doi.org/10.1016/j.jvcir.2013.11.003⟩.

APPLEGATE, D. L.; BIXBY, R. E.; CHVATAL, V.; COOK, W. J. *The Traveling Salesman Problem: A Computational Study (Princeton Series in Applied Mathematics).* Princeton, NJ, USA: Princeton University Press, 2007. ISBN 0691129932, 9780691129938.

BALBONI, A.; MARCHETTI, M.; COLAJANNI, M.; MELEGARI, A. Supporting sense-making and decision-making through time evolution analysis of open sources. In: *2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace.* [S.l.: s.n.], 2015. p. 185–202. ISSN 2325-5374.

BERNERS-LEE, T.; MASINTER, L.; MCCAHILL, M. *Uniform Resource Locators (URL).* [S.l.], 1994. ⟨http://www.rfc-editor.org/rfc/rfc1738.txt⟩. Available at: ⟨http://www.rfc-editor.org/rfc/rfc1738.txt⟩.

BORONDO, J.; MORALES, A. J.; LOSADA, J. C.; BENITO, R. M. Characterizing and modeling an electoral campaign in the context of twitter: 2011 spanish presidential election as a case study. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, v. 22, n. 2, p. 023138, 2012. Available at: ⟨https://doi.org/10.1063/1.4729139⟩.

BORONDO, J.; MORALES, A. J.; LOSADA, J. C.; BENITO, R. M. Analyzing the usage of social media during spanish presidential electoral campaigns. In: . IEEE, 2016. p. 785–792. ISBN 978-1-5090-2846-7. Available at: ⟨http://ieeexplore.ieee.org/document/7752327/⟩.

BOX, G.; JENKINS, G.; REINSEL, G.; LJUNG, G. *Time Series Analysis: Forecasting and Control.* Wiley, 2015. (Wiley Series in Probability and Statistics). ISBN 978-1-118-67492-5. Available at: ⟨https://books.google.com.br/books?id=rNt5CgAAQBAJ⟩.

BREUER, A.; EILAT, R.; WEINSBERG, U. Friend or faux: Graph-based early detection of fake accounts on social networks. In: *Proceedings of The Web Conference 2020.* New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 1287–1297. ISBN 9781450370233. Available at: ⟨https://doi.org/10.1145/3366423.3380204⟩.

CALDARELLI, G.; CHESSA, A.; PAMMOLLI, F.; POMPA, G.; PULIGA, M.; RICCABONI, M.; RIOTTA, G. A multi-level geographical study of italian political elections from twitter data. *PLOS ONE*, Public Library of Science, v. 9, n. 5, p. 1–11, 05 2014. Available at: ⟨https://doi.org/10.1371/journal.pone.0095809⟩.

CAO, Q.; SHEN, H.; GAO, J.; WEI, B.; CHENG, X. Popularity prediction on social platforms with coupled graph neural networks. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020. (WSDM '20), p. 70–78. ISBN 9781450368223. Available at: ⟨https://doi.org/10.1145/3336191.3371834⟩.

CASTEIGTS, A.; FLOCCHINI, P.; QUATTROCIOCCHI, W.; SANTORO, N. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, Taylor & Francis, v. 27, n. 5, p. 387–408, 2012. Available at: ⟨https://doi.org/10.1080/17445760.2012.668546⟩.

CHATFIELD, C. *The Analysis of Time Series: An Introduction*. [S.l.]: CRC Press LLC, 2004. 333 p. ISBN 1584883170.

CHOI, J. J.; HAUSER, S.; KOPECKY, K. J. Does the stock market predict real activity? time series evidence from the g-7 countries. *Journal of Banking and Finance*, v. 23, n. 12, p. 1771 – 1792, 1999. ISSN 0378-4266.

CLAUSET, A.; EAGLE, N. Persistence and periodicity in a dynamic proximity network. In: *2007 Proceedings of the DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*. [S.l.: s.n.], 2007.

CORMEN, T.; LEISERSON, C.; RIVEST, R.; STEIN, C. *Introduction to Algorithms*. MIT Press, 2009. (Computer science). ISBN 9780262033848. Available at: ⟨https://books.google.com.br/books?id=i-bUBQAAQBAJ⟩.

CRABTREE, B. F.; RAY, S. C.; SCHMIDT, P. M.; O'CONNOR, P. T.; SCHMIDT, D. D. The individual over time: Time series applications in health care research. *Journal of Clinical Epidemiology*, v. 43, n. 3, p. 241 – 260, 1990. ISSN 0895-4356. Available at: ⟨http://www.sciencedirect.com/science/article/pii/089543569090005A⟩.

ERTEN, C.; HARDING, P. J.; KOBOUROV, S. G.; WAMPLER, K.; YEE, G. V. Exploring the computing literature using temporal graph visualization. In: *Visualization and Data Analysis*. [S.l.: s.n.], 2004.

FEARNSIDE, P. M. Forests and global warmingnext term mitigation in Brazil: opportunities in the brazilian forest sector for responses to previous termglobal warmingnext term under the clean development mechanism. *Biomass and Bioenergy*, v. 16, n. 3, p. 171 – 189, 1999. ISSN 0961-9534. Available at: ⟨http://www.sciencedirect.com/science/article/B6V22-3VY0BKK-1/2/021355b5a972e8f6fd3996a990f7019f⟩.

FLANIGAN, F.; FRANK, D.; KAZDAN, J.; FRISTEDT, B.; GRAY, L. *Calculus Two: Linear and Nonlinear Functions*. Springer New York, 1998. (Undergraduate Texts in Mathematics). ISBN 9780387973883. Available at: ⟨https://books.google.com.br/books?id=i-p4ovfT4JsC⟩.

GAMA, J. a.; ŽLIOBAITĖ, I.; BIFET, A.; PECHENIZKIY, M.; BOUCHACHIA, A. A survey on concept drift adaptation. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 46, n. 4, mar. 2014. ISSN 0360-0300. Available at: ⟨https://doi.org/10.1145/2523813⟩.

GODSIL, C.; ROYLE, G. F. *Algebraic Graph Theory*. [S.l.]: Springer, 2001. (Graduate Texts in Mathematics, Book 207). ISBN 9781461301639 1461301637.

GUHATHAKURTA, K.; BHATTACHARYA, B.; CHOWDHURY, A. R. Using recurrence plot analysis to distinguish between endogenous and exogenous stock market crashes. *Physica A: Statistical Mechanics and its Applications*, v. 389, n. 9, p. 1874–1882, 2010. ISSN 0378-4371.

HAKIMI, S. L. Optimum locations of switching centers and the absolute centers and medians of a graph. *Oper. Res.*, INFORMS, Institute for Operations Research and the Management Sciences (INFORMS), Linthicum, Maryland, USA, v. 12, n. 3, p. 450–459, jun. 1964. ISSN 0030-364X. Available at: ⟨http://dx.doi.org/10.1287/opre.12.3.450⟩.

HAKIMI, S. L. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, v. 13, n. 3, p. 462–475, 1965. Available at: ⟨https://doi.org/10.1287/opre.13.3.462⟩.

HERMIDA, A.; HERNÁNDEZ-SANTAOLALLA, V. Twitter and video activism as tools for counter-surveillance: the case of social protests in spain. *Information, Communication & Society*, Routledge, v. 21, n. 3, p. 416–433, 2018. Available at: ⟨https://doi.org/10.1080/1369118X.2017.1284880⟩.

HOOTSUITE AND WE ARE SOCIAL. *Digital 2020: Brazil*. Singapore, 2020. Available at: ⟨https://datareportal.com/reports/digital-2020-brazil⟩.

HOSOKAWA, Y.; SHIRATO, H.; NISHIOKA, T.; TSUCHIYA, K.; CHANG, T.-C.; KAGEI, K.; OHOMORI, K.; OBINATA, K.; KANEKO, M.; MIYASAKA, K.; NAKAMURA, M. Effect of treatment time on outcome of radiotherapy for oral tongue carcinoma. *International Journal of Radiation Oncology, Biology, and Physics*, v. 57, n. 1, p. 71–78, 2003. ISSN 0360-3016. Available at: ⟨http://www.sciencedirect.com/science/article/B6T7X-496NMSJ-G/2/12c3224c56818edb9b9b178f7159917f⟩.

JUANG, W.; HUANG, S.; HUANG, F.; CHENG, P.; WANN, S. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in southern taiwan. *BMJ Open*, British Medical Journal Publishing Group, v. 7, n. 11, 2017. ISSN 2044-6055. Available at: ⟨https://bmjopen.bmj.com/content/7/11/e018628⟩.

KÄRNER, O. Arima representation for daily solar irradiance and surface air temperature time series. *Journal of Atmospheric and Solar-Terrestrial Physics*, v. 71, n. 8-9, p. 841 – 847, 2009. ISSN 1364-6826. Available at: ⟨http://www.sciencedirect.com/science/article/B6VHB-4VYXMTN-4/2/afeffae66b68ffb3975800221f1edb88⟩.

KATRAGADDA, S.; VIRANI, S.; BENTON, R.; RAGHAVAN, V. Detection of event onset using Twitter. In: . IEEE, 2016. p. 1539–1546. ISBN 978-1-5090-0620-5. Available at: ⟨http://ieeexplore.ieee.org/document/7727381/⟩.

KIM, D.; OH, H. Emd: A package for empirical mode decomposition and hilbert spectrum. *R Journal*, v. 1, 01 2009.

KIM, D.; OH, H. *EMD: Empirical Mode Decomposition and Hilbert Spectral Analysis.* [S.l.], 2018. R package version 1.5.8.

KIM, H.; ANDERSON, R. Temporal node centrality in complex networks. *Phys. Rev. E*, American Physical Society, v. 85, p. 026107, Feb 2012. Available at: ⟨https://link.aps.org/doi/10.1103/PhysRevE.85.026107⟩.

KO, M. K.; SZE, N. D.; MOLNAR, G.; PRATHER, M. J. Global warming from chlorofluorocarbons and their alternatives: Time scales of chemistry and climate. *Atmospheric Environment. Part A. General Topics*, v. 27, n. 4, p. 581–587, 1993. ISSN 0960-1686. Available at: ⟨http://www.sciencedirect.com/science/article/B757D-48CGB7J-NN/2/5f81c2d3acb62b3f65cb6d71b6152688⟩.

KOÇAK, K.; SAYLAN, L.; EITZINGER, J. Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding. *Ecological Modelling*, v. 173, n. 1, p. 1 – 7, 2004. ISSN 0304-3800. Available at: ⟨http://www.sciencedirect.com/science/article/B6VBS-4BRCMDP-2/2/8e009e2ee3460e0509632d67d0762e25⟩.

KUMAR, N.; JHA, G. K. Time series ann approach for weather forecasting. In: . [S.l.: s.n.], 2013.

LEBARON, B.; ARTHUR, W. B.; PALMER, R. Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control*, v. 23, n. 9-10, p. 1487 – 1516, 1999. ISSN 0165-1889.

LI, L.; WU, Y.; ZHANG, Y.; ZHAO, T. Time+user dual attention based sentiment prediction for multiple social network texts with time series. *IEEE Access*, v. 7, p. 17644–17653, 2019.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.* [S.l.: s.n.], 2002.

LOVINS, J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, v. 11, n. 1-2, p. 22–31, 1968.

LUBBERS, M. J.; VERDERY, A. M.; MOLINA, J. L. Social networks and transnational social fields: A review of quantitative and mixed-methods approaches. *International Migration Review*, v. 54, n. 1, p. 177–204, 2020. Available at: ⟨https://doi.org/10.1177/0197918318812343⟩.

LUHN, H. P. Key word-in-context index for technical literature (kwic index). *American Documentation*, v. 11, n. 4, p. 288–295, 1960. Available at: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090110403⟩.

LUO, X.; JIANG, C.; WANG, W.; XU, Y.; WANG, J.; ZHAO, W. User behavior prediction in social networks using weighted extreme learning machine with distribution optimization. *Future Generation Computer Systems*, v. 93, p. 1023 – 1035, 2019. ISSN 0167-739X. Available at: ⟨http://www.sciencedirect.com/science/article/pii/S0167739X17307938⟩.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. *Natural Language Engineering*, Cambridge university press, v. 16, n. 1, p. 100–103, 2010.

MELLO, R. F. de; RIOS, R. A.; PAGLIOSA, P. A.; LOPES, C. S. Concept drift detection on social network data using cross-recurrence quantification analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, v. 28, n. 8, p. 085719, 2018. Available at: ⟨https://doi.org/10.1063/1.5024241⟩.

MORETTIN, P.; TOLOI, C. de C. *Análise de séries temporais*. Edgard Blucher, 2006. (ABE - Projeto Fisher). ISBN 978-85-212-0389-6. Available at: ⟨https://books.google.com.br/books?id=Q7bJAAAACAAJ⟩.

NGO, C.; MA, Y.; ZHANG, H. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 15, p. 296–305, 2005.

NICOSIA, V.; TANG, J.; MASCOLO, C.; MUSOLESI, M.; RUSSO, G.; LATORA, V. Graph metrics for temporal networks. In: *Temporal Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 15–40. ISBN 978-3-642-36461-7. Available at: ⟨https://doi.org/10.1007/978-3-642-36461-7_2⟩.

NUNES, M.; VIEIRA, F.; ZAVAGLIA, C.; SOSSOLOTE, C.; HERNANDEZ, J. The design of a lexicon for brazilian portuguese: Lessons learned and perspectives. In: *the Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese*. [S.l.: s.n.], 1996. p. 61–70.

PAN, R. K.; SARAMÄKI, J. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, APS, v. 84, n. 1, p. 016105, 2011.

PEW RESEARCH CENTER. *Social Media Use in 2018*. Washington, D.C., 2018.

PEW RESEARCH CENTER. *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Washington, D.C., 2019.

PONOMARENKO, V.; PROKHOROV, M.; BESPYATOV, A.; BODROV, M.; GRIDNEV, V. Deriving main rhythms of the human cardiovascular system from the heartbeat time series and detecting their synchronization. *Chaos, Solitons & Fractals*, v. 23, n. 4, p. 1429–1438, 2005. ISSN 0960-0779. Available at: ⟨http://www.sciencedirect.com/science/article/B6TJ4-4D34JMV-6/2/b5156afe320195f7b8857503cf82fe4a⟩.

PORTER, M. F. *Snowball: A language for stemming algorithms*. 2001. Published online. Accessed 11.03.2008, 15.00h. Available at: ⟨https://snowballstem.org/texts/introduction.html⟩.

RAIESDANA, S.; GOLPAYEGANI, S. M. R. H.; FIROOZABADI, S. M. P.; HABIBABADI, J. M. On the discrimination of patho-physiological states in epilepsy by means of dynamical measures. *Computers in Biology and Medicine*, Pergamon Press, Inc., Elmsford, NY, EUA, v. 39, n. 12, p. 1073–1082, 2009. ISSN 0010-4825.

RIOS, R. A. *Improving time series modeling by decomposing and analysing stochastic and deterministic influences*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 10 2013.

RIOS, R. A.; LOPES, C. S.; SIKANSI, F. H. G.; PAGLIOSA, P. A.; MELLO, R. F. de. Analyzing the public opinion on the brazilian political and corruption issues. In: *2017 Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2017. p. 13–18.

RIOS, R. A.; PAGLIOSA, P. A.; ISHII, R. P.; MELLO, R. F. de. Tsviz: A data stream architecture to online collect, analyze, and visualize tweets. In: *Proceedings of the Symposium on Applied Computing*. New York, NY, USA: ACM, 2017. (SAC '17), p. 1031–1036. ISBN 978-1-4503-4486-9. Available at: ⟨http://doi.acm.org/10.1145/3019612.3019811⟩.

SANTOS, V. M. G. dos; MELLO, R. F. de; NOGUEIRA, T.; RIOS, R. A. Quantifying temporal novelty in social networks using time-varying graphs and concept drift detection. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 650–664. ISBN 978-3-030-61380-8.

SCHLIMMER, J. C.; GRANGER JR., R. H. Incremental learning from noisy data. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 1, n. 3, p. 317–354, mar. 1986. ISSN 0885-6125. Available at: ⟨https://doi.org/10.1023/A:1022810614389⟩.

SHUMWAY, R. H.; STOFFER, D. S. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. 2^a. ed. Springer, 2006. Hardcover. ISBN 0387293175. Available at: ⟨http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387293175⟩.

STEPHEN, M.; GU, C.; YANG, H. Visibility graph based time series analysis. *PLOS ONE*, Public Library of Science, v. 10, n. 11, p. 1–19, 11 2015. Available at: ⟨https://doi.org/10.1371/journal.pone.0143015⟩.

SUBBIAN, K.; PRAKASH, B. A.; ADAMIC, L. Detecting large reshare cascades in social networks. In: *Proceedings of the 26th International Conference on World Wide Web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017. (WWW '17), p. 597–605. ISBN 9781450349130. Available at: ⟨https://doi.org/10.1145/3038912.3052718⟩.

SUMMA, M. G.; STEYAERT, J.; VAUTRAIN, F.; WEITKUNAT, R. A new clustering method for time series to discover geographical cancer trends from 1960 to 2000. *Annals of Epidemiology*, v. 17, n. 9, p. 744–744, 2007. ISSN 1047-2797. Available at: ⟨http://www.sciencedirect.com/science/article/B6T44-4PGJ2CX-27/2/0b189f506dcc57515df516dc8933d8a1⟩.

TANG, J.; MUSOLESI, M.; MASCOLO, C.; LATORA, V. Characterising temporal distance and reachability in mobile and online social networks. *SIGCOMM Comput. Commun. Rev.*, ACM, New York, NY, USA, v. 40, n. 1, p. 118–124, jan. 2010. ISSN 0146-4833. Available at: ⟨http://doi.acm.org/10.1145/1672308.1672329⟩.

TANG, J.; SCELLATO, S.; MUSOLESI, M.; MASCOLO, C.; LATORA, V. Small-world behavior in time-varying graphs. *Phys. Rev. E*, American Physical Society, v. 81, p. 055101, May 2010. Available at: ⟨https://link.aps.org/doi/10.1103/PhysRevE.81.055101⟩.

TANG, X.; XIA, L.; LIAO, Y.; LIU, W.; PENG, Y.; GAO, T.; ZENG, Y. New approach to epileptic diagnosis using visibility graph of high-frequency signal. *Clinical EEG and Neuroscience*, v. 44, n. 2, p. 150–156, 2013. PMID: 23508995. Available at: ⟨https://doi.org/10.1177/1550059412464449⟩.

TIRYAKIOGLU, F.; ERZURUM, F. Use of social networks as an education tool. *Contemporary educational technology*, v. 2, n. 2, 2011.

TSCHACHER, W.; KUPPER, Z. Time series models of symptoms in schizophrenia. *Psychiatry Research*, v. 113, n. 1-2, p. 127–137, 2002. ISSN 0165-1781. Available at: ⟨http://www.sciencedirect.com/science/article/B6TBV-47BXBC4-F/2/fe201ad0573e6f210b9deb2747a98ead⟩.

TSYMBAL, A. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, Citeseer, v. 106, n. 2, p. 58, 2004.

WANG, Y.; YUAN, Y.; MA, Y.; WANG, G. Time-dependent graphs: Definitions, applications, and algorithms. *Data Science and Engineering*, p. 1–15, 09 2019.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in nlp. In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. (COLING '92), p. 1106–1110. Available at: ⟨https://doi.org/10.3115/992424.992434⟩.

WEST, D. *Introduction to Graph Theory*. Pearson, 2017. (Math Classics). ISBN 9780131437371. Available at: ⟨https://books.google.com.br/books?id=61gtAAAACAAJ⟩.

WIDMER, G.; KUBAT, M. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, v. 23, n. 1, p. 69–101, Apr 1996. ISSN 1573-0565. Available at: ⟨https://doi.org/10.1007/BF00116900⟩.

WILSON, R. *Introduction to Graph Theory*. Longman, 1996. ISBN 9780582249936. Available at: ⟨https://books.google.com.br/books?id=tSolAQAAIAAJ⟩.

YANG, L.; LI, Z.; GIUA, A. Containment of rumor spread in complex social networks. *Information Sciences*, v. 506, p. 113 – 130, 2020. ISSN 0020-0255. Available at: ⟨http://www.sciencedirect.com/science/article/pii/S0020025519306607⟩.

YU, Q.; ERHARDT, E. B.; SUI, J.; DU, Y.; HE, H.; HJELM, D.; CETIN, M. S.; RACHAKONDA, S.; MILLER, R. L.; PEARLSON, G.; CALHOUN, V. D. Assessing dynamic brain graphs of time-varying connectivity in fmri data: Application to healthy controls and patients with schizophrenia. *NeuroImage*, v. 107, p. 345 – 355, 2015. ISSN 1053-8119. Available at: ⟨http://www.sciencedirect.com/science/article/pii/S105381191401012X⟩.

YU, S.; CLARK, O. G.; LEONARD, J. J. A statistical method for the analysis of nonlinear temperature time series from compost. *Bioresource Technology*, v. 99, n. 6, p. 1886–1895, 2008. ISSN 0960-8524. Available at: ⟨http://www.sciencedirect.com/science/article/B6V24-4R2XCK3-1/2/0bda6ffef29b835f2c00e8ead28b023c⟩.

ZACHARY, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, [University of New Mexico, University of Chicago Press], v. 33, n. 4, p. 452–473, 1977. ISSN 00917710. Available at: ⟨http://www.jstor.org/stable/3629752⟩.

ZHANG, J.; CENTOLA, D. Social networks and health: New developments in diffusion, online and offline. *Annual Review of Sociology*, v. 45, n. 1, p. 91–109, 2019. Available at: ⟨https://doi.org/10.1146/annurev-soc-073117-041421⟩.

ZHANG, Y.; TANG, J.; SUN, J.; CHEN, Y.; RAO, J. Moodcast: Emotion prediction via dynamic continuous factor graph model. In: *2010 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2010. p. 1193–1198.

ZHAO, R.; ZHOU, A.; MAO, K. Automatic detection of cyberbullying on social networks based on bullying features. In: *Proceedings of the 17th International Conference on Distributed Computing and Networking*. New York, NY, USA: Association for Computing Machinery, 2016. (ICDCN '16). ISBN 9781450340328. Available at: ⟨https://doi.org/10.1145/2833312.2849567⟩.

ZHU, L.; GUO, D.; YIN, J.; STEEG, G. V.; GALSTYAN, A. Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering*, v. 28, n. 10, p. 2765–2777, 2016.

ZHUANG, J. J.; NING, X. B.; HE, A. J.; ZOU, M.; SUN, B.; WU, X. H. Alteration in scaling behavior of short-term heartbeat time series for professional shooting athletes from rest to exercise. *Physica A: Statistical Mechanics and its Applications*, v. 387, n. 26, p. 6553–6557, 2008. ISSN 0378-4371. Available at: ⟨http://www.sciencedirect.com/science/article/B6TVG-4T72WV7-5/2/16aed014159b1067aa6d9b1769371620⟩.

ZWOLENSKI, M.; WEATHERILL, L. The digital universe: Rich data and the increasing value of the internet of things. *Australian Journal of Telecommunications and the Digital Economy*, v. 2, p. 47, 2014.