An Analysis of Political Parties Cohesion based on Congressional Speeches

Willian P. C. Lima[1][0000−0001−5281−9892], Lucas C. Marques[1], Laura S. Assis[1][0000−0003−3081−9722], and Douglas O. Cardoso[2][0000−0002−1932−334X]

[1] Celso Suckow da Fonseca Federal Center of Technological Education, Rio de Janeiro, RJ, Brazil
{willian.lima, lucas.custodio}@aluno.cefet-rj.br, laura.assis@cefet-rj.br
[2] Smart Cities Research Center, Polytechnic Institute of Tomar, Tomar, Portugal
douglas.cardoso@ipt.pt

Abstract. Speeching is an intrinsic part of the work of parliamentarians, as they expose facts as well as their points of view and opinions on several subjects. This article details the analysis of relations between members of the lower house of the National Congress of Brazil during the term of office between 2011 and 2015 according to transcriptions of their house speeches. In order to accomplish this goal, Natural Language Processing and Machine Learning were used to assess pairwise relationships between members of the congress which were then observed from the perspective of Complex Networks. Node clustering was used to evaluate multiple speech-based measures of distance between each pair of political peers, as well as the resulting cohesion of their political parties. Experimental results showed that one of the proposed measures, based on aggregating similarities between each pair of speeches, is superior to a previously established alternative of considering concatenations of these elements relative to each individual when targeting to group parliamentarians organically.

Keywords: Politics · Social Networks · Natural Language Processing · Machine Learning

1 Introduction

Recent research works were based on the relational foundations of politics, giving rise to different models and methodologies to explore such a subject [3, 6, 7, 14, 32]. In this context, a familiar scenario is that of a National Congress of any country wherein debates happen as an inherent aspect of the nation's legislative process. Our work proposes a methodology for knowledge discovery based on the affinity between members of the congress estimated from the speeches given in those debates. Characterizing interactions between parliamentarians is an interesting activity for democracy, as it may provide evidence of suspicious associations and conflicts of interest. Targeting such challenges, we focused on the lower house of the bicameral National Congress of Brazil, known as the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

Chamber of Deputies, which is more dynamic and nuanced than those of some other countries [17].

For this task, the concept of complex networks can be used [4], since it is a very common way of representing relationships between individuals [13]. A network can be defined as a graph in which there is a set of nodes representing the objects under study, and a set of edges connecting these nodes. The edges represent an existing relationship between two nodes according to the context in which they are inserted [24]. In this work, a complete graph was considered in which each node represents a deputy, and the weights associated with the edges that connect each pair of nodes represent the similarities between their political positions which were orally stated.

From the creation of the network using deputies and their speeches, it is possible to analyze the cohesion of the parties to which they belong, as members of ideological groups tend to have similar discourses [8, 11, 15, 25]. The fact that each deputy explicitly belongs to a single political party allows a more objective assessment of the relationship between congressional peers. This can be based on verbal statements of their positions and ideas. Therefore, party cohesion can be estimated according to the panorama across the network of deputies, through quantitative aggregation of the conformity of their speeches. This examination is performed considering speeches of deputies belonging to the same party and that of deputies from different parties. This targeted not only to evaluate the association of deputies to parties but also to highlight phenomena as secessions within parties.

Multiple Natural Language Processing (NLP) approaches were considered to quantify the weights of the edges in the deputies network, based on the cosine similarity between the *TF-IDF (term frequency–inverse document frequency)* vectors regarding their speeches. Our main contributions are: *(i)* establishing measures of distance between deputies based on their speeches; *(ii)* use Agglomerative Hierarchical Clustering to estimate the quality of such measures; *(iii)* evaluate party cohesion from a complex networks perspective.

The remainder of this paper is organized as described next. Section 2 examines the theoretical base that underlies this work. Section 3 includes works related to the problem of measuring textual similarity. The proposed methodology is described in Section 4. The results obtained through the computational experiments performed and their respective discussion are discussed in Section 5. Finally, Section 6 presents conclusions and future work.

2 Theoretical Reference

This section presents the necessary information to understand the developed research properly. First, some concepts of Natural Language Processing are introduced, which serve as a basis for the proposed methods. Subsequently, graph and complex network concepts are presented, and a validation method of clustering is reported.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

2.1 Natural Language Processing

Natural Language Processing (NLP) is the branch of Artificial Intelligence that allows machines to interpret human language combining computational linguistics with statistical models, allowing computers to process human language. NLP uses various pre-processing and text representation techniques aimed at computation to realize automatic text cognition [9].

To model the data and enable an artificial intelligence algorithm a better understand the text and make better associations, it is necessary to perform a pre-processing that abstract and structure the language, ideally preserving only what is the relevant information. The following pre-processing techniques were used in this work [23]: tokenization, stemming, and stop words removal.

A commonly used method to represent the text is to generate a matrix, in which each row is a vector that refers to one of the documents under analysis and each column is relative to a token. A token is a sequence of contiguous characters that play a certain role in a written language, such as words or even parts of them. Each matrix entry is calculated using a widely used measure called *TF-IDF* [5]. It is a manner of determining the piece of content quality based on an established expectation of what an in-depth piece of content contains. This is a statistical measure that is intended to indicate the importance of a word in a document in relation with a collection of documents or a linguistic corpus. The purpose of using *TF-IDF* is to reduce the impact of tokens that occur in a large majority or in very few documents, being of little use to differentiate them. The formula used to calculate the *TF-IDF* for a term t of a document d in a set of documents is given by Equation (1), where $\mathrm{tf}(t, d)$ is the frequency of the term t in the document d.

$$\mathrm{TFIDF}(t, d) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t) \tag{1}$$

The $\mathrm{idf}(t)$ is calculated as shown in Equation (2), where n is the total number of documents, in the document set e $\mathrm{df}(t)$ is the frequency of documents where t occurs. Document frequency is the number of documents in the dataset that contain the term t.

$$\mathrm{idf}(t) = \log\left(\frac{n}{\mathrm{df}(t)}\right) + 1 \tag{2}$$

For two elements of any vector space, their dot product is proportional to the cosine of the angle between them. Such a value can be interpreted as an indication of similarity between these vectors on a scale that varies, if they are unitary, from -1 (diametrically opposite) to 1 (coincident). Using the matrix generated by *TF-IDF*, we can use cosine similarity to measure the similarity between pairs of documents, a measure that can vary from 0 to 1 given that all entries are non-negative. For the calculation of distances between documents, the similarity complement is used. Since a and b are distinct vectors of the matrix

DOI: 10.1007/978-3-031-36024-4_8

TF-IDF the distance between these vectors is given by Equation (3):

$$d(a, b) = 1 - (\langle a, b \rangle)/(||a|| \cdot ||b||) \quad (3)$$

2.2 Clustering

There are several ways to perform a clustering, and one of them is through hierarchical agglomerative clustering methods [31]: initially each group contains a single element, and iteratively, the two most similar groups are concatenated until, in the end, the number of groups is equal to the number of parties. To define the clustering criterion of the agglomerative cluster, the hyper-parameter linkage is used. The linkage criterion determines which distance to consider between the observation sets. The algorithm will merge the cluster pairs considering the minimization of this criterion. The following types of linkage were considered in this work: average, complete, and single.

When organizing the collection items into clusters, there are two criteria that can be used to evaluate the resulting cluster: *i)* homogeneity (h) and *ii)* completeness (c). Both depend on the establishment *a priori* of a cluster considered ideal for the data, indicating the respective class for each item. A clusterization has a maximum homogeneity value if all its clusters contain only samples that are originally members of the same class, without merging elements that should be separate. Completeness, on the other hand, is maximum if all samples that are members of a given class are elements of the same cluster, and have not been improperly allocated to different groups.

In some ways, cluster homogeneity and completeness are competing. The trivial clustering, in which each item is in a unitary cluster, has maximum homogeneity but minimum completeness, while this scenario is reversed if all items are placed in a single group. The *V-Measure* [27] measures the success of the criteria of homogeneity and completeness in a combined way, as can be seen in Equation Equation (4):

$$V_\beta = ((1 + \beta)hc)/(\beta h + c) \quad (4)$$

The *V-Measure* is a measure calculated as the harmonic mean of the values of homogeneity and completeness. These values can be adjusted to assign different weights to the contributions of homogeneity or completeness through the parameter $\beta \geq 0$. Usually, a default value of $\beta = 1$ is used. If $\beta \in [0, 1[$, the homogeneity has greater relevance in the weighting. However, if $\beta \in (1, \infty)$ the completeness has greater weight in the calculation. In the case of $\beta = 1$, there is a balance between the two criteria. The closer the result of V-Measure is to 1, the better the performance.

A measure called *silhouette coefficient* can be used to validate the consistency of the clustering process. It is calculated for each item, based on the average distance of it to other items which belong to the same cluster (a) and the minimum average distance of it to elements of another cluster (b), as shown in Equation (5). The closer to 1 is the average silhouette, the better was the clustering process.

$$silhouette = \frac{(b-a)}{max(a,b)} \tag{5}$$

In complex networks, the clustering coefficient measures the degree to which graph nodes tend to cluster. In its most basic definition, considering a given node u, the clustering coefficient represents the relative frequency of triangles formed by u and its neighbors. There are several manners to perform such an evaluation for weighted graphs [28]. One of them is based on the geometric mean of the edges weights of each triangle, as described in Equation (6).

$$ca_u = \frac{\sum\limits_{v,z \in N^u} (\hat{w}_{uv}.\hat{w}_{uz}.\hat{w}_{vz})^{1/3}}{deg(u).(deg(u)-1))}, \tag{6}$$

where $deg(u)$ is the degree of node u. Nodes v and z are neighbors of u, thus N^u is the neighborhood of the node u. $\hat{w}_{ab}$ is the weight of the edge (a,b). Evidence suggests that in most real-world networks, and especially in social networks, nodes tend to form strongly connected groups characterized by a relatively high density of loops. This probability inclines to be greater than the average probability of a loop being randomly established between two nodes [29].

3 Related Works

Historically, political scientists have devoted much attention to the role that political institutions and actors play in a variety of phenomena in this context. Recently, however, there has been a strong trend towards a point of view based on the relational foundations of politics, giving rise to various political networks and methodologies to characterize their structures [20].

Some of the most recent works related to this topic use different information to associate parliamentarians with each other. For example, voting, campaign donations, and participation in events. In [7] complex networks are used to assess the relationship between donations received by parliamentarians elected in 2014 and their voting behavior during 2015 and 2016. The authors examine the homophily and cohesion of parliamentarians in networks created concerning their political parties and constituencies.

Determining the thematic profile of federal deputies, through the processing of texts obtained from their speeches and propositions is held in [16]. This work presents natural language processing techniques used to analyze the speeches of deputies. Such speeches include the removal of words with little semantic meaning (stop words), terms reduction into their morphological roots (stemming), computational representation of texts (bag-of-words), and uses of the Naive Bayes model for the classification of discourses and propositions.

The challenge of relating the deputies, by the similarity of their votes, using the votes in which the parliamentarians participate was addressed in [6].

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

This approach can be applied in a variety of political scenarios, as most current legislative processes have votes for the approval of the proposals.

A network technique approach to relating different portals of news according to the ideological bias of the published news is presented in [2]. The paper uses hyperlinks to develop an automatic classification method in order to analyze the relationship between the network's structural characteristics and political bias. That is if the ideology of portals is reflected in the properties of the networks that model citations (hyperlinks) among them. In the works [33] and [12] NLP and machine learning techniques were used to determine news bias. They use overtly political texts for a fully automatic assessment of the political tendency of online news sites.

Another major area of research is those focused on social network analysis, as approached by [15], which used tweets to automatically verify the political and socio-economic orientation of Chilean news portals. Sentiment analysis techniques were applied in [11] to assess the relationship between the opinion of Twitter users (text in Portuguese) and the elections final result. The same occurs for other countries and languages [25], considering studies on identification of political positioning and ideology through textual data.

Although a greater abundance of works on NLP are focused on supervised learning, there are also those in the literature on this topic aimed at unsupervised learning. For example, there are works [13, 30] on methods of grouping textual objects, where they are automatically organized into similar groups and a comparative study of grouping algorithms applied to the clustering of such objects is carried out. In [18], a modeling of topics of speeches in the European Union parliament is carried out.

4 Methodology

In this section, we describe a baseline approach to the proposed study based on previous works, the proposed method in this work to overcome the presented previous literature, and the measure used to evaluate the considered approaches.

4.1 Preliminary Approach

To define the distance between parliamentarians based on their speeches, the User-as-Document (UaD) [10] model was used. This consists of concatenating all the speeches of a parliamentarian, thus forming a single document that represents him. Then, the similarity between two parliamentarians is calculated using the similarity between their respective documents.

To create these documents, two NLP pre-processing techniques are used: *(i)* removal of stop-words and other tokens considered less relevant, and *(ii)* stemming [22]. The list of stop-words was based on the list available in Portuguese in the package NLTK [21] with the addition of very common words in the political context, such as the sra , v. exa among others. For the execution of stemming, the module of NLTK SnowballStemmer was applied to the corpus. After this

DOI: 10.1007/978-3-031-36024-4_8

step, the documents are represented in a vector way using *TF-IDF* and the similarities between them are calculated by the cosine similarity.

An illustration that exemplifies a network where the nodes represent parliamentarians, the edges are determined by the similarity between them according to their discourses is presented in Figure 1. The vectors that goes along with each vertex represent the speeches of each parliamentarian. The speeches which are classified by subject/idea through the colors, and the values are the number of speeches given of each type. The calculation of similarities is nothing more than the dot product of these vectors. It is carried out by multiplying the number of speeches of a deputy by that of another deputy. This multiplication occurs only among the number of the same type discourses, that is, situated in the same coloring. The sum of these multiplications results in the weight value of each edge.

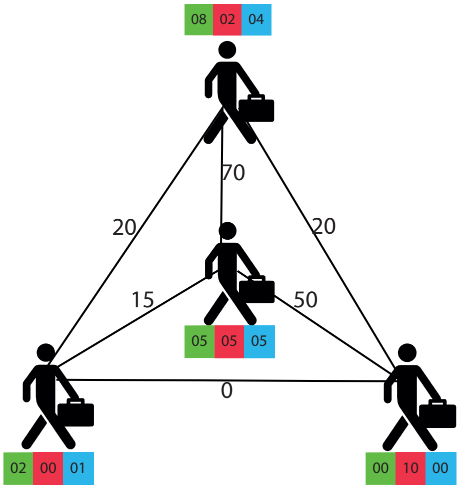

Fig. 1: Illustration of the similarity evaluation between parliamentarians through their speeches.

4.2 Proposed Approach

In the context of parliamentarians' speeches, the use of the UaD approach can cause some difficulties in calculating distances. One reason is the mischaracterization of the number of documents in the corpus, which directly reflects on the *TF-IDF* calculation. Another related problem is the terms crossing from different discourses, which can lead to an exaggerated extrapolation in the parliamentarian characterization.

An alternative approach to assessing the alignment between two parliamentarians is to calculate the distance between each pair of their speeches, and then aggregate that collection of distances into a single value that represents the distance between parliamentarians. In this way, the distribution of tokens in the generated corpus would be more realistic compared to the one corresponding achieved from the UaD approach, employing the NLP techniques used in the preliminary approach. Furthermore, the impact of crossing different terms between discourses can be minimized using different aggregation methods.

It is possible to apply several aggregation methods to summarize in a single measure the distances between the pairs of speeches of any two parliamentarians. In this work, six types of aggregation are proposed. All of them are premised on the formation of a complete bipartite graph determined in such a way that each vertex corresponds to a discourse, and each partition of vertices corresponds to a parliamentarian. Each edge weight represents the distance between the discourses on which it falls. The aggregation types are described next:

1. **Average of distances**: It uses the average of all values distances between pairs of discourses, which corresponds to the edges' average weight of the bipartite graph.
2. **Minimum distances**: It refers to the smallest value of the distance between each pair of discourses, which corresponds to the smallest of the edge weights.
3. **Maximum distances**: It refers to the largest value of the distance between each pair of discourses, which corresponds to the largest of the edge weights.
4. **Average of shortest distances (AverageMin)**: Measure relative to the average of the smallest distances of each speech.
5. **Average of longest distances (AverageMax)**: Measure relative to the average of the greatest distances of each speech.
6. **Minimal spanning tree (MinST)**: The sum of minimum spanning tree edges of the bipartite graph.

4.3 Validation of Results

In order to legitimize the use of graphs and NLP to determine organic groups of parliamentarians, the use of an objective evaluation criterion was idealized. Thus, it would be possible to identify the optimal configuration of the aggregation type and the pre-processing hyper-parameters. This criterion would ideally indicate how well a clustering reflects the prevailing party structure. However, it should be noted that this does not preclude such clustering from including evidence that contrasts with the aforementioned structure, due to the similarity (or dissimilarity) between parliamentarians.

In this work, the clusters of parliamentarians are defined according to their two-by-two distances, and the number of groups is defined by the number of parties among the parliamentarians, resulting in 21 clusters. Given these conditions, an agglomerative hierarchical clustering approach was used, using predefined similarities instead of coordinates of points whose distance could be evaluated [19].

For the construction of a cluster applying agglomerative hierarchical clustering, three types of criteria applicable to the algorithm to define the clusters

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

were used in this work: *(i) single linkage*, *(ii) average linkage* and *(iii) complete linkage*. From the agglomerative hierarchical clustering, it is possible to assess the quality of clustering, by comparing the membership of parliamentarians to clusters with the membership of parliamentarians to parties. To obtain the clustering quality assessment value, the V-Measure was used. Figure 2 presents a flowchart indicating the steps performed to obtain the evaluation measure.

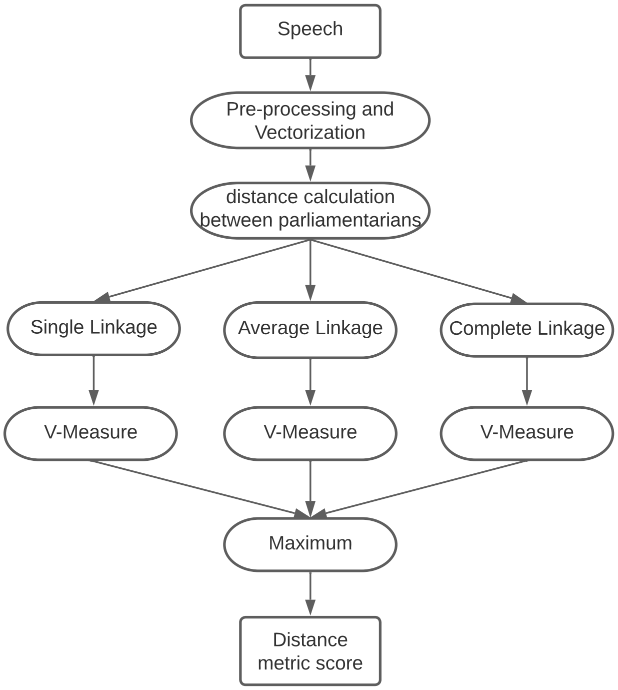

Fig. 2: Flowchart with each step taken to obtain the quality assessment measure of the clusters generated from the applied methodology.

5 Experimental Evaluation

In this section, experiments related to the questions raised in the previous section are detailed. The results of these experiments were obtained using the $8,378$ speeches of the 627 active federal deputies during the 54^a House legislature, which spans the years from 2011 to 2015. Each congressman was treated as a party member to which he belonged when he was elected, disregarding possible party changes during the term of office. The parliamentary data used in this work are open data, available in the Chamber of Deputies Open Data Service [1]. All experiments were performed on the Google Colaboratory platform free of charge.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

5.1 Hyper-parameter Optimization and Correlation Analysis

Hyper-parameters fine-tuning can help achieving to simultaneously reflect the current party structure as well as evidence that contrasts with that structure. Using the combinations of hyper-parameters for vectorization based on the *TF-IDF* statistic and the agglomerative hierarchical clustering algorithm, different versions of the method to measure the distance were used for comparison. In order to obtain the results of all possible configurations in a set of values specified for each hyper-parameter, a grid search was performed [26], with 120 value settings considering, in addition to linkage, the following dimensions, for the *TF-IDF* method:

- **max_df** (MD): When building the vocabulary, terms with a document frequency greater than the given threshold are ignored. The values considered for this hyper-parameter were: $2^0, 2^{-1}, \cdots, 2^{-5}$;
- **max_features** (MF): Builds a vocabulary by selecting no more than this number of attributes, ordered by their frequencies throughout the corpus. The values considered for this hyper-parameter were: $10^1, 10^2, \cdots, 10^5$;
- **sublinear_tf** (ST): Replaces tf (term frequency) with $1 + log(tf)$ in the *TF-IDF* calculation. The values considered for this hyper-parameter were: *Yes* and *No*;
- **use_idf** (UI): Enables the use of inverse-document-frequency (idf) in the vectorization process, which can also be performed considering only *tf*. The values considered for this hyper-parameter were: *Yes* and *No*.

Table 1 contains the best results among all hyper-parameter configurations, according to the highest values of V-Measure, which add *i)* homogeneity and *ii)* completeness. In the context of this experiment, the first represents the minimization of party diversity within each group of parliamentarians inferred based on the similarities between their discourses. The second represents the preservation of party unity in the obtained grouping, so that the fragmentation of originally related parliamentarians is avoided. Most of the results presented use the distance aggregation approach proposed here (Section 4.2): the best configuration used the Average of distances; only one of the top seven is based on the User-as-Document approach.

Looking at the hyper-parameters, approaches, and aggregation methods individually, it is possible to see that the option $max\_df = 2^{-4}$ generally has the best performance among the evaluated values. This can be interpreted as an indication that there are many relatively frequent and non-discriminating terms in the analyzed corpus. The max_features parameter for the method that calculates the distance between each pair of deputies obtains better results with the maximum amount of attributes $\geq 10^4$, while the User-as-Document approach obtained your best result using a maximum of 10^3 attributes. In general, better results were achieved using *True* as an argument for both sublinear_tf and use_idf parameters.

To finalize the comparison of distance measures between parliamentarians, Figure 3 presents the correlation matrix of the considered measures. This matrix is colored like a heat map for better visualization. Kendall's τ was used

Table 1: Best hyperparameter settings found according to the highest values of V-Measure.

Average	**V-measure**	**Linkage**	**max_df**	**max_features**	**sublinear_tf**	**use_idf**
Average	0,2264	average	2^{-4}	10^5	No	Yes
UaD	0,2156	average	2^{-4}	10^3	No	Yes
AverageMax	0,2015	complete	2^{-3}	10^5	Yes	Yes
MinST	0,1950	complete	2^{-4}	10^4	Yes	No
AverageMin	0,1659	average	2^{-4}	10^4	Yes	Yes
Maximum	0,1635	complete	2^0	10^5	Yes	Yes
Minimum	0,1598	complete	2^{-4}	10^2	Yes	No

as correlation statistic, i.e a non-parametric correlation measure, that is able to capture even non-linear relationships between random variables, making the analysis more flexible and robust. According to the presented matrix, the segregation of the UaD measure from the others is evident, also it can be observed the Average is positioned as an intermediary between this first and the remaining five measures. Such an organization is consistent with the definition of the assessed measures. However, attention is drawn to the fact that the strongest correlation is between the alternatives AGMin and AverageMax, despite the two having principles that do not appear to be similar.

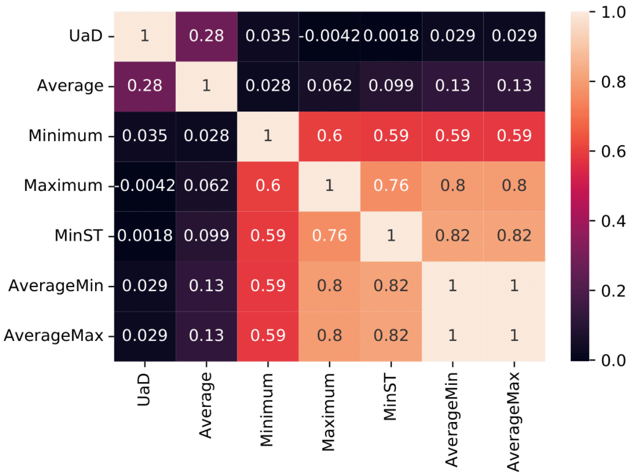

Fig. 3: Correlation matrix between the measures used to estimate the similarity between parliamentarians.

5.2 Party Cohesion Analysis

Figure 4 depicts party cohesion based on the parliamentary distance measure in its already detailed optimal configuration (Avarage). The order of the parties is arranged from left to right from the highest to the lowest clustering coefficient of the subgraph induced by the vertices relative to parliamentarians of each party. On the ordinate axis, the values of the clustering coefficient are shown together with the party acronym. The clustering coefficient evaluates the degree to which the graph vertices tend to colligate, considering the distance between the vertices of the same group. Parties whose parliamentarians have speeches at small distances from each other will have larger clustering coefficients. Smaller clustering coefficients indicate greater distances between speeches by parliamentarians from a given party. The red bars represent the party size in terms of the parliamentarians' number, while the blue bars characterize the entropy present in the distribution of parliamentarians of a party in clusters. An entropy value equal to zero means that all the members of a party were clustered into the same cluster. As the members of a party are dispersed into more groups the entropy increases.

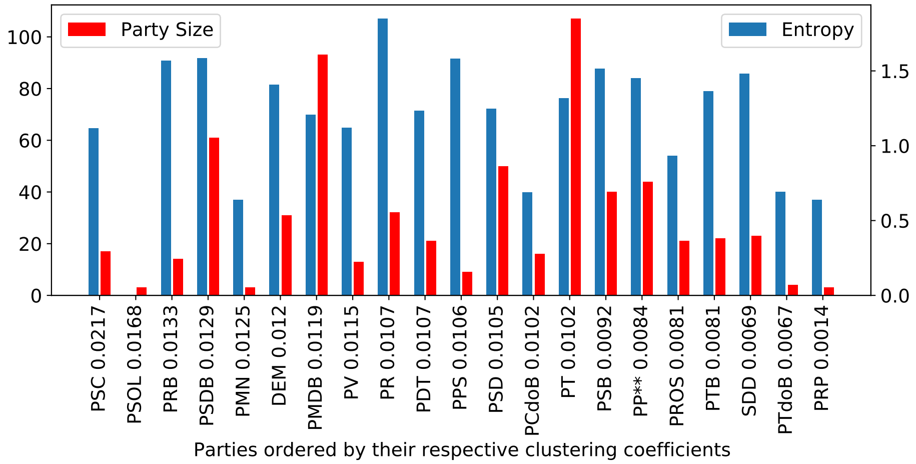

Fig. 4: Assessment of similarity between parliamentarians.

From Figure 4 it is possible to analyze the relationship between party size, clustering coefficient and entropy based on the graph of distances between parliamentarians. It is expected that parties with greater numbers of parliamentarians have higher entropy and lower clustering coefficient, resulting in less party cohesion, as there is a greater chance of divergences between parliamentarians, given the large number. It is possible to observe that there are cases where this expectation is confirmed, as in the second party with the best clustering coefficient, the PSOL, whose number of parliamentarians is small, and its entropy is all small and clustering coefficient is larger. Other examples of parties that also follow this expectation are the PT, PSB and PP, where the number of parliamentarians is large, and entropy is also large and clustering coefficient is lower. Some parties do not correspond to this behavior, such as the PRP and the PTdoB, which have a very small number of parliamentarians and a relatively high entropy and lower clustering coefficient. These results suggest great distances between the speeches of parliamentarians that belong to the same party.

The PMDB and PSDB, parties with large numbers of parliamentarians, are examples of parties with a relatively high value of clustering coefficient, and with a high value of entropy, indicating small distances between parliamentarians of a single party, and also a dispersion of parliamentarians from one party to other parties. This means that the process of clustering spread these parties' parliamentarians into several other clusters, despite the similarity between the speeches of their parliamentarians on average. This can be observed as evidence of the existence of cohesive and distant wings in these parties.

6 Conclusion

Analyzing the relations between deputies and parties is an interesting activity for democracy, because it provides data about deputies and possible conflicts of interest. In this work, the relationship and party cohesion of parliamentarians of the Deputies Chamber active during the 54^a legislature, which comprises the years between 2011 and 2015, according to their speeches, was analyzed. Complex networks were used to model the relationships between deputies. An unsupervised machine learning approach, named agglomerative hierarchical clustering, was used for the analysis of party cohesion.

The proposed methodology assesses the affinity between any pair of parliamentarians based on the discourses sets of each of them. For this, a complete bipartite graph was considered. In this graph, the vertices represent the parliamentarians' speeches, divided into two partitions according to the parliamentarian to which they correspond. The edges are weighted according to the respective speeches' textual similarity. Affinity can then be calculated by aggregating these weights according to one of the six alternatives presented and considered for this purpose. The presented methodology obtained superior results compared to the reference measure previously established in the literature. In the meantime, the aggregation method that obtained the best results was the one related to the average of the similarities of each pair of speeches , considering any two parliamentarians.

In the future, we intend to use other Natural Language Processing techniques to obtain different measures in addition to *TF-IDF*. In the next study, we want to consider word embedding, since *TF-IDF* is a simpler technique that may not consider the nuances of language as synonyms, for example. We also intend to test different methods to compute the measures of aggregation using algorithms in graphs. Such algorithms can lead to measures that better reflect the distances between parliamentarians based on their speeches.

Acknowledgements Douglas O. Cardoso acknowledges the financial support by the Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia, FCT) through grant UIDB/05567/2020, and by the European Social Fund and programs Centro 2020 and Portugal 2020 through project CENTRO-04-3559-FSE-000158.

References

1. Dados abertos da câmara dos deputados. https://dadosabertos.camara.leg.br/swagger/api.html, (Accessed on 08/17/2021)
2. Aires, V., da Silva, A., Nakamura, F., Nakamura, E.: An evaluation of structural characteristics of networks to identify media bias in news portals. In: Proceedings of the Brazilian Symposium on Multimedia and the Web. pp. 225–232 (2020)
3. Aref, S., Neal, Z.P.: Identifying hidden coalitions in the US House of Representatives by optimally partitioning signed networks based on generalized balance. Scientific Reports **11**(1), 19939 (Oct 2021)
4. Barabási, A.L., Pósfai, M.: Network science. Cambridge University Press, Cambridge, United Kingdom (2016), oCLC: ocn910772793
5. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2016)
6. Brito, A.C.M., Silva, F.N., Amancio, D.R.: A complex network approach to political analysis: Application to the brazilian chamber of deputies. Plos one **15**(3), e0229928 (2020)
7. Bursztyn, V.S., Nunes, M.G., Figueiredo, D.R.: How brazilian congressmen connect: homophily and cohesion in voting and donation networks. Journal of Complex Networks **8**(1), cnaa006 (2020)
8. Caetano, J.A., Lima, H.S., Santos, M.F., Marques-Neto, H.T.: Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 american presidential election. Journal of internet services and applications **9**(1), 1–15 (2018)
9. Camacho-Collados, J., Pilehvar, M.T.: On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. arXiv preprint arXiv:1707.01780 (2017)
10. Cossu, J.V., Labatut, V., Dugué, N.: A review of features for the discrimination of twitter users: Application to the prediction of offline influence. Social Network Analysis and Mining **6**(1), 25 (2016)
11. Cristiani, A., Lieira, D., Camargo, H.: A sentiment analysis of brazilian elections tweets. In: Anais do VIII Symposium on Knowledge Discovery, Mining and Learning. pp. 153–160. SBC (2020)
12. Dallmann, A., Lemmerich, F., Zoller, D., Hotho, A.: Media bias in german online newspapers. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media. pp. 133–137 (2015)
13. Dias, M., Braz, P., Bezerra, E., Goldschmidt, R.: Contextual Information Based Community Detection in Attributed Heterogeneous Networks. IEEE Latin America Transactions **17**(02), 236–244 (Feb 2019)
14. Doan, T.M., Gulla, J.A.: A Survey on Political Viewpoints Identification. Online Social Networks and Media **30**, 100208 (Jul 2022)
15. Elejalde, E., Ferres, L., Herder, E.: The nature of real and perceived bias in chilean media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. pp. 95–104 (2017)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8

16. Fernandes, M.S.: Tenho dito: uma aplicação para análise de discursos parlamentares utilizando técnicas de processamento de linguagem natural (2017)
17. Gomes Ferreira, C.H., Murai Ferreira, F., de Souza Matos, B., Marques de Almeida, J.: Modeling Dynamic Ideological Behavior in Political Networks (2019)
18. Greene, D., Cross, J.P.: Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Political Analysis **25**(1), 77–94 (2017)
19. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3), 264–323 (1999)
20. Lee, S.H., Magallanes, J.M., Porter, M.A.: Time-dependent community structure in legislation cosponsorship networks in the congress of the republic of peru. Journal of Complex Networks **5**(1), 127–144 (2016)
21. Loper, E., Bird, S.: Nltk: The natural language toolkit. CoRR **cs.CL/0205028** (2002)
22. Lovins, J.B.: Development of a stemming algorithm. Mech. Transl. Comput. Linguistics **11**(1-2), 22–31 (1968)
23. Méndez, J.R., Iglesias, E.L., Fdez-Riverola, F., Díaz, F., Corchado, J.M.: Tokenising, stemming and stopword removal on anti-spam filtering domain. In: Conference of the Spanish Association for Artificial Intelligence. pp. 449–458. Springer (2005)
24. Metz, J., Calvo, R., Seno, E.R., Romero, R.A.F., Liang, Z., et al.: Redes complexas: conceitos e aplicações. (2007)
25. Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S.L., Celdrán, A.H., Pérez, M.G., Ruipérez-Valiente, J.A., Pérez, G.M., Mármol, F.G.: Spotting political social bots in twitter: A use case of the 2019 spanish general election. IEEE Transactions on Network and Service Management **17**(4), 2156–2170 (2020)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp. 410–420 (2007)
28. Saramäki, J., Kivelä, M., Onnela, J.P., Kaski, K., Kertesz, J.: Generalizations of the clustering coefficient to weighted complex networks. Physical Review E **75**(2), 027105 (2007)
29. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature **393**(6684), 440–442 (1998)
30. Wives, L.K.: Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de" clustering" (1999)
31. Yim, O., Ramdeen, K.T.: Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. The Quantitative Methods for Psychology **11**(1), 8–21 (Feb 2015)
32. Zhang, X., Wang, H., Yu, J., Chen, C., Wang, X., Zhang, W.: Polarity-based graph neural network for sign prediction in signed bipartite graphs. World Wide Web **25**(2), 471–487 (Mar 2022)
33. Zhitomirsky-Geffet, M., David, E., Koppel, M., Uzan, H.: Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites. Online Information Review (2016)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_8