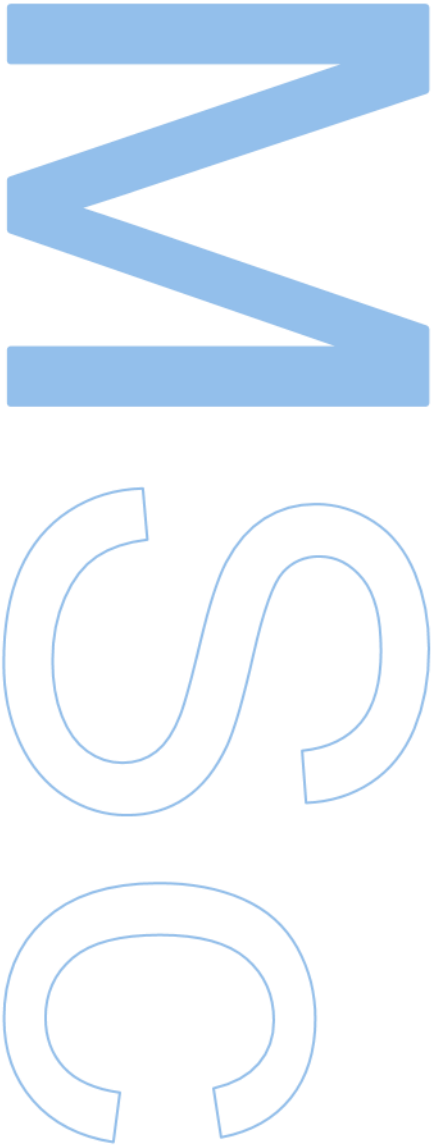
Recommender Systems for vote prediction

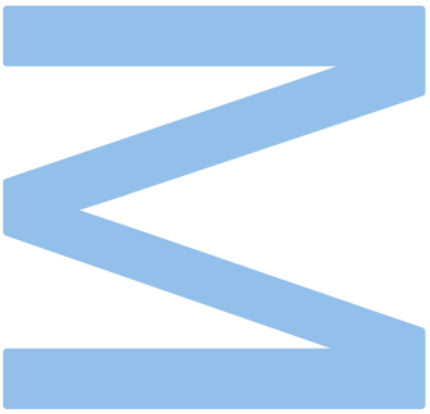


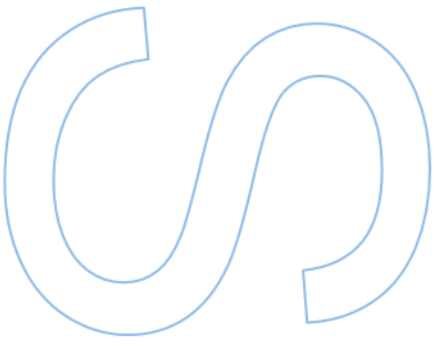
Rita Lima Cardoso
Data Science
Departamento de Ciência de Computadores
2023

Supervisor
Prof. Dr. João Vinagre, Professor Auxiliar Convidado, Faculdade de Ciências

Co-supervisor
Prof. Dr. Douglas Cardoso, Investigador, Instituto Politécnico de Tomar


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





DECLARAÇÃO DE HONRA

Eu, **Rita Lima Cardoso**, inscrito(a) no Mestrado em **Data Science** da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Rita Lima Cardoso

[Porto, 24 de Agosto de 2023]

Abstract

This master's thesis focuses on the development of a recommendation system tailored specifically for voting scenarios. Voting is a fundamental aspect of democratic processes, and making informed decisions is crucial for active citizen participation. Recommendation systems have become integral to various domains, including e-commerce, entertainment, and information dissemination. However, the lack of transparency in these systems raises concerns about privacy, fairness, and algorithmic biases. The proposed recommendation system analyzes historical voting patterns to generate personalized voting recommendations for government personnel. Through this research, we contribute to the field of recommendation systems by focusing on voting patterns within democratic houses. By harnessing the power of recommendation algorithms, we strive to optimize voting behavior, promote transparency, and improve decision-making processes within political parties and government bodies.

Keywords: Data analysis; Recommendation algorithms; Voting scenarios; Electoral data; Government agencies; Voting behavior.

Resumo

Esta dissertação de mestrado concentra-se no desenvolvimento de um sistema de recomendação adaptado especificamente para cenários de votação. A votação é um aspeto fundamental dos processos democráticos, sendo crucial tomar decisões informadas para uma participação cívica ativa. Os sistemas de recomendação tornaram-se indispensáveis em vários domínios, incluindo comércio eletrónico, entretenimento e disseminação de informação. No entanto, a falta de transparência nesses sistemas suscita preocupações relacionadas com privacidade, equidade e possíveis distorções algorítmicas. O sistema de recomendação proposto analisa padrões históricos de votação para gerar recomendações personalizadas de voto para o pessoal governamental. Através desta pesquisa, contribuímos para o campo dos sistemas de recomendação, focando nos padrões de voto dentro de entidades governamentais. Aproveitando o poder dos algoritmos de recomendação, procuramos otimizar o comportamento de voto, promover a transparência e melhorar os processos de tomada de decisão nos partidos políticos e nos organismos governamentais.

Palavras-Chave: Análise de dados; Algoritmos de recomendação; Cenários de votação; Dados eleitorais; Órgãos governamentais; Comportamento de voto.

Agradecimentos

A jornada desta tese tem sido marcada por inúmeros desafios, conquistas ao longo do ano, tristezas, alegrias e incertezas. No entanto, mesmo sendo um caminho solitário, não o concluo sozinha.

Não posso deixar de expressar a minha gratidão a ambos os meus orientadores, Prof.João Vinagre e Prof.Douglas Cardoso. Por mostrarem o apoio e os insights necessarios ao meu trabalho, por serem extremamente pacientes e nunca me deixarem de qualquer forma mal.

Gostaria de estender os meus sinceros agradecimentos ao meu grupo de amigos, que sempre me proporcionou encorajamento para prosseguir, pela inspiração que constantemente me transmitem e pelo apoio moral desde o primeiro dia. Ao meu namorado, agradeço pelo o apoio incondicional, partilha e companheirismo, e por nunca ter duvidado das minhas capacidades.

E, é claro, não posso deixar de expressar o meu profundo reconhecimento àqueles que me proporcionaram tudo isto, a minha família. O vosso apoio constante e inabalável tem sido a maior força por trás de todas as minhas conquistas.

Em suma, desejo reconhecer todos aqueles que contribuíram direta ou indiretamente para a conclusão bem-sucedida da minha tese. O vosso apoio e confiança nas minhas capacidades exerceram um papel indispensável, e por isso, estou imensamente agradecida.

Contents

List of Tables

List of Figures

Acronyms

CF Collaborative Filtering

DP Democratic Party

MAE Mean Absolute Error

PL Partido Liberal

PP Partido Popular

PSL Partido Social Liberal

RMSE Root-mean-square error

RP Republican Party

RS Recommendation Systems

USA United States of America

Chapter 1

Introduction

Recommendation systems (RS) are algorithms designed to suggest consumers with similar and pertinent content, aiming to increase user engagement and satisfaction. These systems have gained significant importance with the rise of the Web as a transnational platform. The primary goal of RS is to provide users with customized recommendations that align with their interests and preferences. By considering users' past interactions, RS aims to introduce them to relevant products and increase product sales, ultimately generating profit. Through personalized recommendations, RS enhances the user experience by presenting chosen items tailored to individual preferences, leading to increased engagement and user satisfaction.

E-commerce is one of the most well-known applications of recommendation algorithms. Businesses like Amazon use these algorithms to make product recommendations to their customers, based on their past shopping and browsing history (Smith and Linden, 2017). The entertainment industry also extensively uses RS algorithms, especially on streaming services like Netflix, Disney+, and Spotify. These types of platforms' recommendations for new content are based on analyzing users' viewing and listening habits to improve user engagement and experience.

Although recommendation systems are frequently used in e-commerce or entertainment, there has been limited exploration of their potential in less obvious fields, like political science. The use of such systems in this domain has the potential to assist individuals in making informed decisions and improving democratic processes.

However, while these algorithms can be very convenient, they are not without restrictions. One of the biggest problems encountered by an RS is data sparsity and, usually, this problem occurs when most users either don't interact with most of the items or the available rating is sparse. Another common problem is when a brand new user joins - it can also happen to a new item - and in this case, it is complicated to compute recommendations, since the user or item in question does not have past interactions yet.

1.1 Motivation

Politics is an activity that depends on a group of individuals to decide on a variety of issues and the allocation of power in a society. It has an impact on how that culture perceives crucial issues like economic, sociological, and political advancement. Making wise and deliberate decisions during political events like voting on laws or elections would therefore be extremely crucial to the democratic society of a particular country.

Pursuing this further, Brazil and the United States of America (USA) are both democratic nations, each with its own distinct political system. Brazil operates under a presidential and federal system of government, where power is divided among the president, the National Congress, and the Supreme Court. The National Congress consists of two houses, the Chamber of Deputies and the Senate. However, Brazil's democracy faces challenges including increasingly polarized politics, a weak central state, and a lack of trust in the government (Sabatine and Wallace, 2023).

Similarly, the USA is a federal presidential republic with three branches of government: executive, legislative, and judicial. The President leads the executive branch, Congress is responsible for making laws in the legislative branch, and the Supreme Court ensures adherence to the Constitution (Kaufman, 2022). The USA also grapples with issues of political polarization and a lack of trust in the government, mirroring the challenges faced by Brazil.

To shed light on the potential of recommendation systems in politics, we conducted a thorough literature review and analyzed case studies. The literature review revealed that big data has been utilized to gather public opinion on political issues. Some notable examples include:

- In the 2016 United States Presidential election, Cambridge Analytica, a political consulting firm, implemented a recommendation system to target political advertising on Facebook by analyzing user data to predict their political inclinations and tailor ads accordingly (Cadwalladr,2018).
- In the 2018 presidential elections in Brazil, the Workers' Party employed data analytics to target specific demographics and influence people's voting behavior (Layton et al.,2021).
- During the 2014 Lok Sabha elections in India, the Bharatiya Janata Party (BJP) utilized data analytics and social media to target specific demographics and influence voting behavior (Safi,2017).

While the assessment of the literature provides insights into the use of recommendation systems in politics, it is crucial to note that future analysis will focus on the potential of recommendation systems in the context of Brazil and the United States. By examining the specific challenges and characteristics of these democratic nations, we aim to explore how recommendation systems can be leveraged to address their unique political landscapes and contribute to informed decision-making processes.

1.2 Main Goals

As evident from Figure 1.1, the initial step toward accomplishing our primary objectives involves analyzing the deputies' data to gain insights into their behavior. Building upon this analysis, our intention is to evaluate recommender systems, thereby examining the voting patterns and inclinations of deputies and parties in both Brazil and the United States.

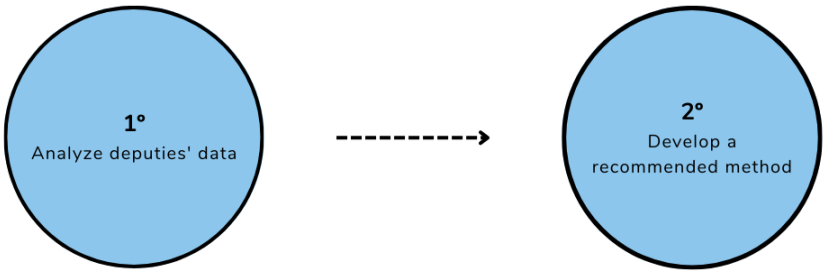


Figure 1.1: Both of the main goals of this dissertation.

By employing recommendation techniques, it becomes possible to evaluate and predict the voting behavior of representatives in these countries. This can contribute to the understanding of political dynamics and decision-making processes in legislative bodies.

1.3 Dissertation Layout

This dissertation will be structured into five comprehensive parts, each addressing specific aspects of the research topic.

In the first chapter 1, the focus will be on setting the context for the project. This includes providing an overview of the research area, highlighting its significance, and stating the main objectives and goals of the dissertation.

The subsequent chapter 2 delves into a comprehensive review of existing literature related to recommendation systems. This chapter serves two purposes: first, it aims to establish a solid foundation of knowledge by examining the existing body of work on recommendation systems.

Second, it analyzes and synthesizes relevant literature that has been influential in shaping the current research.

Chapter 3 focuses on the implementation of the recommendation system and the various approaches employed. This chapter provides a detailed description of the methodology used in developing the system, including the algorithms, data sources, and technical considerations. Additionally, it discusses the different approaches considered and explains their relevance to the research objectives.

Chapter 4 presents the experiments and tests conducted to evaluate the recommendation system's performance. The chapter starts by introducing the evaluation metrics and methodology used to assess the system's effectiveness. It then delves into the generation of realistic test data, which simulates real-world scenarios for robust evaluation. The chapter provides insights into the experimental setup and methodology adopted to ensure accurate and reliable evaluation of the recommendation system.

Chapter 5 is dedicated to the discussion and analysis of the results obtained from the implementation of the recommendation system. It presents a comprehensive evaluation of the system's performance, including metrics, comparisons, and insights derived from the collected data. The chapter provides a critical assessment of the outcomes and interprets them in the context of the research goals.

Ultimately, the final chapter, chapter 6, summarises the dissertation by reviewing the key findings, considering their implications, and addressing any limits or prospective pathways for further research. This chapter further reiterates the project's importance and contributions to the field of recommendation systems.

Chapter 2

State of the art

2.1 Recommendation Systems

A recommender system is an artificial intelligence algorithm designed to provide personalized recommendations to users by predicting their potential interests. It analyzes and leverages their past actions, interactions, or preferences. Its main goal is to correctly forecast a user's likelihood of interest in or happiness with a specific item, even if the person is unfamiliar with it beforehand. The analysis and exploitation of the user's prior behaviors, interactions, or preferences are used to make this prediction.

Recommendation systems have a wide range of applications, including e-commerce, music streaming platforms, and video streaming services. According to MacKenzie et al., recommendations for products account for roughly 35% of Amazon's consumed items, whereas Netflix accounts for approximately 75%.

These systems encompass several types, each characterized by distinct underlying assumptions and methodologies. Within the realm of recommendation systems, we can consider two primary entities: users and items. Users would be the individuals that interact with the systems and provide the due input through actions, such as rating items, purchasing products, or expressing preferences. These actions will serve as indicators of the user's likes and interests.

This type of algorithm operates with two kinds of data:

1. **User-item interaction data**: This includes information about the past actions or behaviors of users, such as ratings given to items, items purchased, or items interacted with. The recommender system can better comprehend the user's tastes and preferences with the use of this data, which offers insights into user preferences.

2. **Information about either user or item**: This includes additional contextual information about users or items that can be used to enhance the recommendation process. For users, this may include demographic information, location, past purchase history, or explicit preferences

provided by the user. This may include attributes such as genre, category, keywords, or descriptions for items. The new information enhances the comprehension of user preferences and item features, resulting in more accurate and pertinent recommendations.

Methods that use the first type of data are usually referred to as collaborative filtering, and the ones that use the latter are called content-based. There is also a type of recommendation system that join both aspects of the mentioned algorithms, entitled a hybrid recommendation system. They combine the strengths of both methods to perform more firmly in a wide assortment of settings. [Aggarwal, 2016]

Due to the unique nature of our challenge, we would perceive the deputies as users and their individual votes as the items in our analysis, treating each voting instance as a discrete unit and focusing on the specifics of how they vote within sessions as an outcome recommendation.

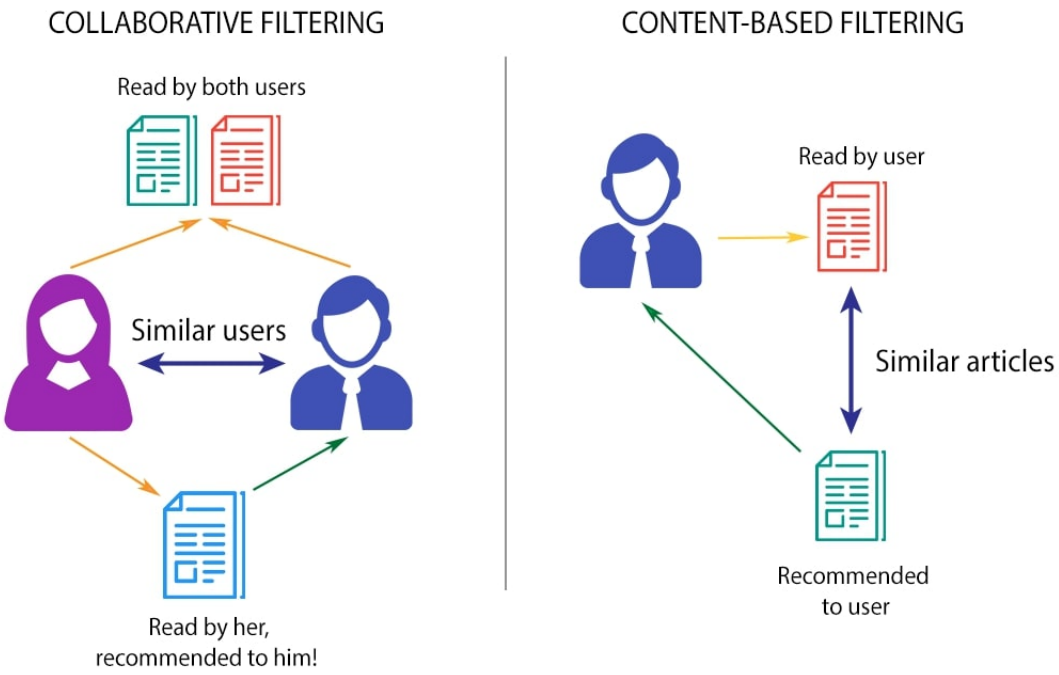


Figure 2.1: Different Methods Used in Recommendation Systems from [9]

As we can see in Figure 2.1, the two principal methods utilize two different techniques. In collaborative filtering, we can use users' past behavior to make recommendations. It looks at the actions of similar users, such as what they've bought or liked, and makes suggestions based on that information. For example, if we were to compare user A and user B, we have two articles that were both read by both of the users and have similar ratings, the system might recommend an article that user A like and reads to user B. Meanwhile, content-based filtering uses information about the items themselves to make recommendations. It looks at the characteristics of items a user has liked or interacted with and suggests other similar items. For example, if the user has read and liked an article about politics, the system will most likely recommend a similar article, with a similar topic.

The objective of both these systems is to construct a user-item matrix, also known as a utility matrix, which consists of the ratings appointed by each user for each item. The matrix is structured with rows representing users and columns representing items. Each cell within the matrix contains the ratings or preferences provided by each user for each item. This representation allows for a clear and organized view of the rating data, making it easier to analyze and extract meaningful insights.

Initially, this matrix is often sparse, as it only contains a limited number of pairs with actual ratings. The ultimate goal of recommendation systems is to identify similarities between users and items, which can be achieved by decomposing the utility matrix into two matrices with the equation 2.1

$$\hat{Y} = UV^T \tag{2.1}$$

where U and V are representations of users and items, respectively, in a low-dimensional space (NS, 2020). This factorization process completes the matrix by uncovering latent features of user-item preferences, thus allowing for the generation of personalized recommendations for users. In our scenario, the matrix would symbolize both deputies and individual voting instances, where each cell of the matrix embodies their respective voting behaviors.

In this illustration, the matrix (Table 2.1) displays user ratings for four films (M1, M2, M3, M4) from three users (U1, U2, U3). The ratings are given on a scale from 1 to 5, with "-" denoting a user who has not submitted a rating for a particular film.

	M1	**M2**	**M3**	**M4**
U1	2	-	3	2
U2	3	4	3	2
U3	2	5	-	2

Table 2.1: Example Matrix of Movie Ratings by Users

This example matrix provides a simplified representation of how users' ratings for movies can be organized in a tabular format. Such a matrix is a common input for recommendation systems, where the goal is to predict or suggest ratings for unseen movies based on the patterns and preferences observed in the existing ratings.

2.1.1 Collaborative Filtering

Collaborative Filtering approaches rely on past interactions between users and items to make predictions. It assumes that individuals with similar tastes and preferences will rate products similarly and that these ratings can be used to predict the ratings of other users for the same objects (Herlocker et al., 2004). Furthermore, collaborative filtering can be separated into model-based and memory-based techniques.

2.1.1.1 Model-Based Method

The model-based method uses supervised and unsupervised machine learning algorithms in order to produce predictions. One commonly used technique is matrix factorization, which decomposes the user-item rating matrix into lower-dimensional latent factors. The model then estimates missing ratings and makes personalized recommendations based on these latent factors. The matrix factorization model can be represented as:

$$R \approx U \cdot V^T \tag{2.2}$$

where R is the user-item rating matrix, U is the user matrix representing user preferences, and V is the item matrix representing item attributes.

2.1.1.2 Memory-Based Method

Memory-based method, also known as a neighborhood-based algorithm, is a technique in which the ratings of user-item combinations are essentially predicted through their neighborhoods. These neighborhoods can be defined by 2 different methods: user-based or item-based collaborative filtering. The user-based method focuses on determining users that have similar tastes to the target user and recommends ratings for the non-classified ratings of that target user by computed weighted averages of the ratings of the determined neighbors. On the other hand, the idea for the latter method consists of first determining a set of items that are the most similar to our target items and using the ratings of those set of items to determine the target item's ratings for the unrated users.

This approach tends to offer high simplicity in implementation, yielding recommendations that are easily understandable. However, there can be a challenge associated with this methodology. For example, if we encounter a sparse rating matrix, this type of method usually will not work well, since it lacks enough data for accurate predictions (Aggarwal,2016).

The memory-based method, also known as a neighborhood-based algorithm, is a technique in which the ratings of user-item combinations are essentially predicted through their neighborhoods. These neighborhoods can be defined by two different methods: user-based or item-based collaborative filtering.

The user-based method focuses on determining users that have similar tastes to the target user and recommends ratings for the non-classified ratings of that target user by computing weighted averages of the ratings of the determined neighbors. Mathematically, the predicted rating (r_{ui}) for a user u and an item i using the user-based collaborative filtering can be calculated as presented in the following equation:

$$r_{ui} = \bar{r}_u + \frac{\sum\limits_{v \in N(u)} \text{sim}(u, v) \cdot (r_{vi} - \bar{r}_v)}{\sum\limits_{v \in N(u)} |\text{sim}(u, v)|} \tag{2.3}$$

where $\bar{r}_u$ is the average rating given by user u, $N(u)$ is the set of users similar to user u, $\text{sim}(u, v)$ represents the similarity between users u and v, r_{vi} is the rating of user v for item i, and $\bar{r}_v$ is the average rating given by user v.

On the other hand, the item-based method involves determining a set of items that are most similar to our target item and using the ratings of those items to determine the target item's ratings for unrated users. The predicted rating (r_{ui}) for a user u and an item i using the item-based collaborative filtering can be calculated as:

$$r_{ui} = \frac{\sum_{j \in N(i)} \text{sim}(i, j) \cdot r_{uj}}{\sum_{j \in N(i)} |\text{sim}(i, j)|} \tag{2.4}$$

where $N(i)$ is the set of items similar to item i and $\text{sim}(i, j)$ represents the similarity between items i and j.

Memory-based methods are generally simple to implement and can provide understandable recommendations. However, they may encounter challenges when dealing with sparse rating matrices, as they require sufficient data for accurate predictions.

These collaborative filtering methods play a significant role in recommendation systems by leveraging the past interactions of users and items to make personalized recommendations.

2.1.2 Content-Based Models

Content-based filtering approaches rely on the attributes of the items and users to make predictions. They use the characteristics of the items a user has interacted with, such as their genre, artist, or keywords, to recommend similar items to the user. These approaches assume that users who have shown interest in certain attributes in the past are likely to be interested in other items with similar attributes.

The recommendation is calculated with cosine similarity. Cosine similarity is a measure of the similarity between two vectors of an inner product space that measures the space between them. The cosine similarity between two non-zero vectors is measured by taking the product of the vectors and dividing it by the product of their magnitudes - Equation 2.5.

Furthermore, in this type of algorithm, cosine similarity is used to compare the preference vector of various users and items. The mentioned vector is a representation of a user's or item's preferences, and it is often formed by taking the dot product of a user-item matrix with a set of latent factors.

$$Cosine(x, y) = \frac{x \cdot y}{|x||y|} \tag{2.5}$$

Using content-based models has the advantage of not requiring information from other users to make recommendations. They rely solely on the attributes of the items and the user's preferences,

making them suitable for situations where user data is limited or unavailable.

Content-based models provide personalized recommendations based on the similarity of item attributes and user preferences. This approach can effectively capture users' individual interests and provide relevant recommendations.

2.1.3 Hybrid Recommendation System

The hybrid recommendation systems make predictions using both content-based and collaborative filtering. Combining both methods, this technique joins the strengths of both and can overcome any drawback of an individual recommendation technique.

The integration of collaborative filtering and content-based filtering techniques in hybrid recommendation systems has garnered significant attention in the field of recommendation algorithms. These systems strive to overcome the limitations inherent in individual approaches by synergistically combining their capabilities. By leveraging collaborative filtering's ability to capture user preferences based on historical interactions and content-based filtering's focus on item attributes, hybrid recommendation systems offer the potential to provide more accurate and personalized recommendations. Such hybridization is particularly valuable in scenarios where data sparsity or the cold start problem hampers the performance of standalone methods. Through a meticulous fusion of collaborative filtering and content-based filtering, hybrid systems hold the promise of delivering enhanced recommendation accuracy, addressing the overspecialization issue, and facilitating the provision of diverse and tailored recommendations to users.

2.2 Case Studies

In this section, we will present an overview of the most recent big data strategies for use in political elections.

These case studies highlight how data analytics, recommendation systems, and targeted advertising are becoming increasingly important in political campaigns around the world. While data-driven systems enable precision targeting and personalized messages, they also raise questions about privacy, data ethics, and the possibility of democratic process manipulation. As political campaigns change, it is more important to strike a balance between harnessing data insights and guaranteeing openness, accountability, and the preservation of individuals' rights.

2.2.1 Cambridge Analytica's Influence in the 2016 US Election

Cambridge Analytica, a political consulting firm, gained significant attention during the 2016 United States Presidential election. Their approach involved utilizing data analytics and a recommendation system to target political advertising on Facebook, with the aim of influencing

voters based on their individual preferences and political inclinations (Cadwalladr,2018). The firm's practices raised concerns about privacy, data ethics, and the impact of personalized messaging on the democratic process.

2.2.2 Data-Driven Campaigning in the 2015 UK General Election

The use of data in political campaigns was evident during the 2015 UK General Election. In Anstead's academic paper titled "Data-driven Campaigning," they explore the increased use of data in political campaigns, including voter targeting, message testing, and digital advertising. The paper discusses the potential benefits and drawbacks of data-driven campaigning, highlighting the need for transparency and ethical considerations in the use of data (Anstead,2017).

The paper highlights several key aspects of data-driven campaigning observed during the 2015 UK General Election. One prominent area is voter targeting, where political parties and candidates leveraged data analytics to identify specific demographic groups and tailor their campaign messages accordingly. By analyzing vast amounts of data, including voter registration records, socio-economic information, and consumer behavior, campaigns could segment the electorate and customize their messages to resonate with different voter groups.

2.2.3 Data Analytics in the 2012 US Presidential Election

The 2012 United States Presidential election showcased various uses of data in political campaigns. The campaigns employed data-driven strategies for micro-targeting voters, analyzing demographic trends, and optimizing campaign messaging. Data analysis and predictive modeling played a significant role in understanding voter behavior and tailoring campaign efforts to specific segments of the population (Hellweg, 2014).

2.2.4 Data Analytics in the 2018 Brazilian Presidential Election

In the 2018 presidential election in Brazil, the Workers' Party utilized data analytics to target specific demographics and influence people's voting behavior. By leveraging data analytics, the party aimed to understand voter preferences, tailor their campaign messages, and strategically engage with potential supporters. This case study demonstrates the growing use of data-driven approaches in political campaigns, raising questions about privacy, ethics, and the implications of targeted messaging (Layton et al., 2021).

2.2.5 Data Analytics and Social Media in the 2014 Indian Lok Sabha Elections

During the 2014 Lok Sabha elections in India, the Bharatiya Janata Party (BJP) employed data analytics and social media to target specific demographics and influence voting behavior.

The party utilized data-driven strategies to understand voter sentiments, identify key issues, and deliver personalized messages to potential voters. This case study illustrates the role of data analytics and social media in political campaigns, emphasizing the impact of targeted communication and the ethical considerations surrounding data usage (Safi, 2017).

2.3 Summary

In this chapter, we have explored the critical role of recommendation systems in enhancing user experiences and driving revenue for businesses. We have discussed the underlying algorithms used in these systems, including collaborative filtering, content-based filtering, and hybrid approaches. By leveraging user-item interactions and item attributes, these algorithms generate personalized recommendations that cater to individual user preferences.

Furthermore, we have examined the construction of user-item matrices and the decomposition into low-dimensional representations of users and items. Model-based methods, such as matrix factorization, and memory-based methods, such as neighborhood-based approaches, have been highlighted as effective techniques for recommendation generation. Additionally, we have discussed content-based models that rely on item attributes and user preferences to make recommendations without requiring data from other users.

Throughout this chapter, we have also presented notable case studies in the realm of recommendation systems, such as the involvement of Cambridge Analytica in the 2016 US Presidential election and the use of data-driven campaigning in the 2015 UK general election. These case studies demonstrate the practical application and impact of recommendation systems in real-world scenarios, particularly in the realm of big data-driven political campaigns.

Chapter 3

Methodology

3.1 Problem Formulation

In the context of our problem, we designate the deputies as users and their individual voting instances as items. Each deputy's profile is analogous to a user, while each separate vote cast by a deputy constitutes an item in our analysis. This unique perspective allows us to tailor our recommender system to capture the intricacies of voting behaviors across various contexts.

For our recommender system, we interpret the concept of "rating" distinctively. Rather than conventional numerical scores, our ratings are defined by the deputies' voting behaviors. These interactions come together to give us a complete picture of how deputies tend to vote.

Our main goal is to uncover meaningful patterns and connections from these interactions. By doing this, we want to gain valuable insights into deputies' voting preferences and provide practical recommendations based on their past voting behaviors.

3.2 Data Collection and Preprocessing

This section provides an overview of the data collection process and outlines the essential steps to preprocess the acquired data before incorporating it into the recommendation system.

Effective data collection and preprocessing are vital to ensure the data's accuracy, reliability, and suitability for subsequent analysis and modeling tasks.

3.2.1 Data Collection

One of the countries we chose to analyze was Brazil, a nation of particular interest for our study. In order to obtain the needed data, we turned to the official website of the da Câmara dos Deputados (also known as Chamber of Deputies), a prominent and authoritative platform in the domain

of Brazilian politics. It served as a comprehensive platform, offering a wide range of legislative information encompassing voting records, bill proposals, and deputy profiles. Recognizing the significance of these sources, they were identified as the primary data repositories for the present study. Specifically, the data for the entire year of 2022 was selected for comprehensive evaluation at the commencement of this research endeavor.

Our analysis also encompassed the United States of America, and for that, we used Voteview, an esteemed platform developed by UCLA's Department of Political Science and Social Science Computing. It provides access to a wealth of valuable data related to political behavior and voting patterns. With its extensive collection of historical voting data and comprehensive measures of ideological positions, it serves as a significant resource for researchers and analysts in the field of political science. Recognizing the reliability and relevance of this dataset, it was identified as a primary source of information for this study.

Our research encompassed two distinct political contexts: Brazil, characterized by a multiparty system, and the United States, known for its bipartite landscape. Through this comparative analysis, we sought to understand how the presence of multiple or two major parties influences the political landscape and policy outcomes in each respective country.

3.2.2 Data Preprocessing

Data preparation plays an important role in every data analysis or machine learning project. It entails putting raw data in a format that will be appropriate for subsequent modeling or analysis (Stedman et al., 2022). To assure the reliability and accuracy of the Brazil and USA datasets, a variety of techniques and methodologies were used for data preprocessing.

Data cleaning is an essential stage in the data preprocessing process to assure the consistency, dependability, and accuracy of the datasets. To deal with inconsistencies, errors, and missing values in the context of the thesis, which focuses on datasets from Brazil and the United States, comprehensive data-cleaning procedures were used.

Dealing with missing numbers was one of the main difficulties in the data cleansing process. We carefully chose to eliminate columns with more than 80% missing data to maintain the analysis's quality. With this strategy, the chance of making mistakes due to significant missing data was minimized and the remaining data was ensured to be sufficiently complete.

Moreover, we identified and addressed a particular data leakage vulnerability in Brazil's dataset. This leakage occurred when information from the target variable was unintentionally included in the features, leading to biased analysis results. To counteract this issue, we implemented a rigorous data cleansing procedure to eliminate any instances of data leakage. Our focus was primarily on the voting records, ensuring that each deputy's vote was associated with a unique vote ID. This measure prevented multiple instances of the same deputy casting the same vote, thereby enhancing the integrity of the dataset.

The Brazil and USA datasets were turned into clean, consistent, and acceptable forms for further analysis or modeling by adopting these preprocessing approaches. Data preprocessing is critical for guaranteeing the correctness and dependability of any future data-based analysis, recommendations, or decision-making processes.

3.2.3 Data Structure

In the context of my thesis dissertation, I undertook a comprehensive analysis of Brazilian legislative data sourced from the Camara dos Deputados website. This research endeavor involved the integration of multiple datasets to culminate in an extensive dataset encompassing pivotal attributes. This dataset serves as a repository of pertinent information, shedding light on intricate parliamentary proceedings and facilitating an in-depth comprehension of legislative dynamics within the Brazilian political context.

Central to this analytical pursuit is the meticulously curated dataset featuring the following key columns:

- **idVotacao**: This column houses the distinctive identifier assigned to each voting event, providing a cardinal reference point for monitoring and categorizing individual voting instances.
- **siglaOrgao**: Representing the acronym associated with the legislative entity where the vote transpired, this column imparts valuable insights into the specific organizational milieu.
- **idEvento**: The unique identifier attributed to each event tied to a voting instance is meticulously captured, enabling a granular tracing of legislative occurrences related to the individual votes.
- **aprovacao**: Emblematic of the voting outcome, this column delineates whether the vote garnered approval (1.0), met disapproval (0.0), or assumed an alternative disposition (2.0).
- **uriVotacao**: Serving as an integral hyperlink, this column contains the URI linking to comprehensive information concerning the voting event on the Camara dos Deputados website, furnishing immediate access to contextual depth.
- **voto**: This column provides insight into the intricate array of voting preferences exhibited by legislators on specific issues. It encompasses a spectrum of six distinct voting outcomes, each carrying its own connotations
- **deputado_id**: This column houses the unique identifier allocated to each legislator, affording a seamless linkage between individual legislators and their corresponding voting activities.
- **deputado_nome**: Conveying the full nomenclature of the legislator, this column facilitates the unequivocal identification of members within the legislative body.

- **deputado_siglaPartido**: Capturing the political party affiliation of legislators, this column contributes indispensable insights into the intricate tapestry of partisan dynamics.
- **orientacao**: Expressing the orientation of the legislator's vote, such as "Não" for dissent or "Sim" for concurrence, this column provides elucidation on the alignment of legislators with their party's ideological position.
- **idLegislaturaFinal**: Emblematic of the ultimate legislature identifier, this column furnishes contextual information pertaining to the specific legislative term concurrent with the voting event.

This meticulously crafted dataset stands as the bedrock of my research endeavor, serving as an indispensable artifact for the exploration of Brazilian legislative behavior, discernment of voting trends, and the nuanced interplay of political affiliations. Through the methodical amalgamation of diverse datasets and the scrupulous construction of data columns, my aim is to unravel the underlying dynamics that underpin the decision-making processes within the Brazilian legislative landscape.

For USA voting data, I obtained a comprehensive dataset from the reputable VoteViews website([14]). This dataset, named 'df_voting', serves as the foundational bedrock of my research, embodying a meticulously designed data structure featuring six pivotal columns, each of which plays a significant role in shaping the analytical discourse.

The **Congress** column stands as a temporal indicator, denoting the unique numerical identifier of the legislative session during which individual voting instances transpired. This temporal dimension serves as a cornerstone for tracking and contextualizing voting trends across various congressional tenures.

Situated within the dataset, the **Chamber** column encapsulates the legislative context. By classifying each vote according to its occurrence in the House, Senate, or even the Presidential domain, this column adds a fundamental contextual layer that enriches the understanding of each voting event.

At the heart of the data structure, the **Rollnumber** column carries a narrative thread. It assigns a distinct reference code to the inaugural roll call vote of each congress or legislative session, serving as a pivotal point of origin for the nuanced exploration of voting patterns and their evolution over time.

The **ICPSR** column, essential for individualization, assigns unique identification codes to legislators. Notably, some legislators have multiple ICPSR codes due to party transitions or presidential tenures, underscoring the complexities of their legislative journeys.

In terms of voting preferences, the **Cast_code** column employs a codified numerical spectrum (ranging from 1 to 9) to encapsulate the essence of legislators' decisions. This standardized coding mechanism crystallizes voting inclinations, facilitating the discernment of voting behaviors and alignments that traverse ideological and partisan boundaries.

The **Prob** column introduces a probabilistic dimension to the dataset, derived through the NOMINATE methodology and anchored in variables from the 'df_members' dataset with dual dimensions. These probabilistic estimates offer a nuanced glimpse into the decision-making realm of legislators, fostering a deeper understanding of the factors guiding their voting choices.

In conclusion, my thesis dissertation delves into two distinct yet interrelated spheres of political analysis: Brazilian legislative data from the Camara dos Deputados website and USA voting data sourced from the esteemed VoteViews website. Through rigorous efforts in data integration and meticulous curation, I have constructed comprehensive datasets that serve as critical foundations for my research journey. In both instances, these datasets embody the pinnacle of a meticulous and systematic approach encompassing data acquisition, intricate transformation, and comprehensive analysis. These artfully curated datasets stand as indispensable artifacts pivotal for unveiling the intricate tapestry of legislative dynamics. Beyond their primary role as repositories of information, these datasets serve as the foundational pillars for the development of innovative data analysis and recommendation systems.

3.3 Data Analysis

The investigation of previously discussed material concerning the voting process in Brazil and the United States required the employment of two independent analyses, each of which provided useful insights into the political landscapes of their respective countries. While the datasets differed in content, they were subjected to distinct analyses that provided equivalent viewpoints on Brazil's and the United States political processes.

For Brazil, in total three datasets related to voting in Brazil in 2022 were utilized, one containing detailed voting information by political party and the other summarizing the voting outcomes. A dataset representing elected deputies in the Chamber of Deputies was also included.

As part of the data analysis, an examination was conducted on the voting frequency of political parties 3.1. The findings indicate that the parties PL, PT, PP, UNIÃO, and PSD received the highest number of votes. Significantly, this distribution highlights the predominant presence of right-wing parties within the legislative landscape under investigation. Visual representations, such as graphs, clearly depict the voting patterns, facilitating a comprehensive understanding of each party's involvement in the decision-making process.

In the da Câmara dos Deputados, there are six different ways of voting, (Nunes Moni da Silva et al., 2018):

- "Yes" - Indicates acceptance of the proposal.
- "No" - Indicates rejection of the proposal.
- "Abstention" - Indicates no opinion and is only counted for quorum purposes.

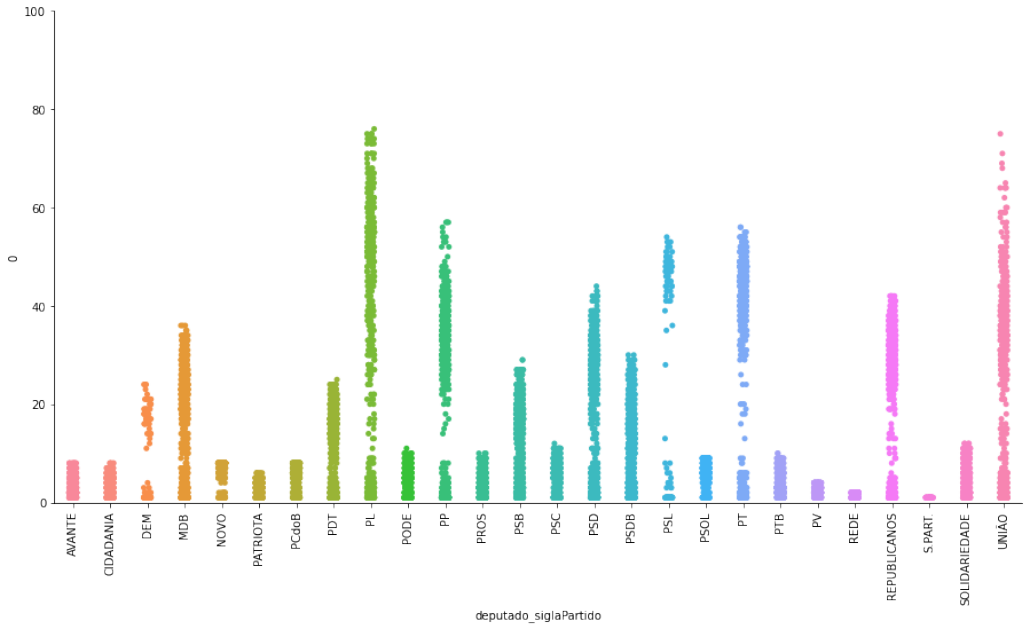


Figure 3.1: Analysis of the Quantity of Votes Received by Each Political Party in the Brazilian Chamber of Deputies

- "Obstruction" - Indicates a decision not to vote or express obstruction. It does not count for quorum purposes.
- "Art. 17" - Represents the value of the neutral vote of the current Chamber's president. The president's vote only holds significance in the event of a tie, where it becomes necessary to break it.
- "Null" - Indicates no available data.

Based on this information, we conducted an analysis to understand the distribution of these voting types in our selected dataset. Our findings, which can be seen in Figure 3.2, revealed that the majority of votes predominantly fall into the "Yes" or "No" categories. Additionally, we observed a limited number of "Art. 17" votes, and an even smaller number of abstentions.

As part of our analysis, we sought to examine any significant shifts in political party affiliations among the deputies 3.3. Our findings revealed that in 2022, over 150 deputies in the Chamber of Deputies in Brazil underwent at least one party change. Among these deputies, we identified instances where an individual had been affiliated with up to four different political parties. Remarkably, this phenomenon occurred three times, involving three distinct deputies, one of whom emerged as the member with the highest number of votes.

Interestingly, these three deputies shared a common association with three specific political parties: PL, PSL, and União. However, the fourth political party varied between S.PART and REPUBLICANOS. It is worth noting that all of these parties fall within the right-wing spectrum of Brazilian politics.

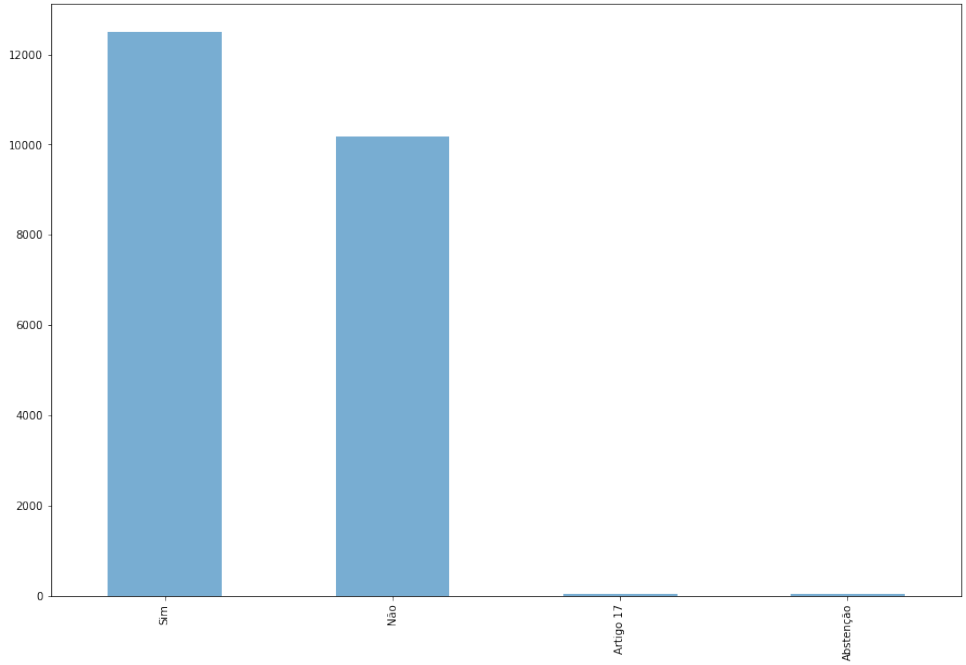


Figure 3.2: Analysis of Vote Types and Distribution in the Brazilian Chamber of Deputies

Figure 3.3: Analysis of Party Switching Among Brazilian Deputies

Continuing our analysis, we turned our attention to the dataset dedicated to the deputies, which provided valuable information about their characteristics, including gender, age, and state of origin. By exploring these variables, we aimed to gain a deeper understanding of the demographic composition within the Chamber of Deputies and identify any notable patterns or trends. Through this examination, we sought to shed light on aspects such as gender representation, age distribution, and regional representation among the deputies.

The analysis of graphics in Figure 3.4 reveals that a mere 3.5 of the Brazilian deputies in the Chamber of Deputies are female, with a total of 272 members. Additionally, it is noteworthy that the primary birthplaces of these deputies are concentrated in Rio de Janeiro (Figure 3.5). Furthermore, the age distribution depicted in the box plot illustrates a wide range, spanning

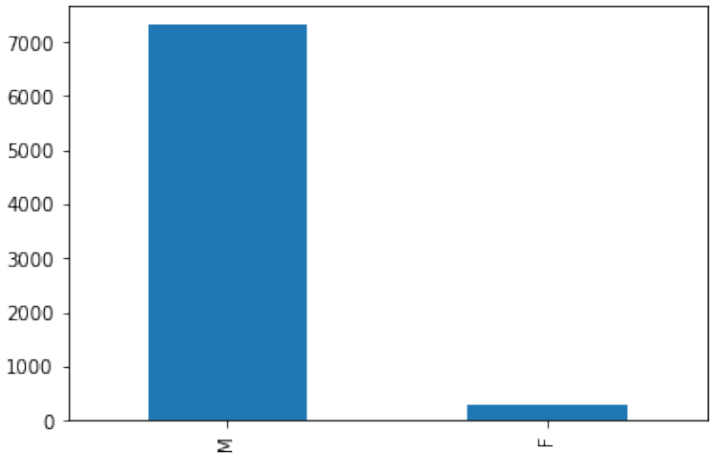


Figure 3.4: Distribution of Gender in the Brazilian Chamber of Deputies

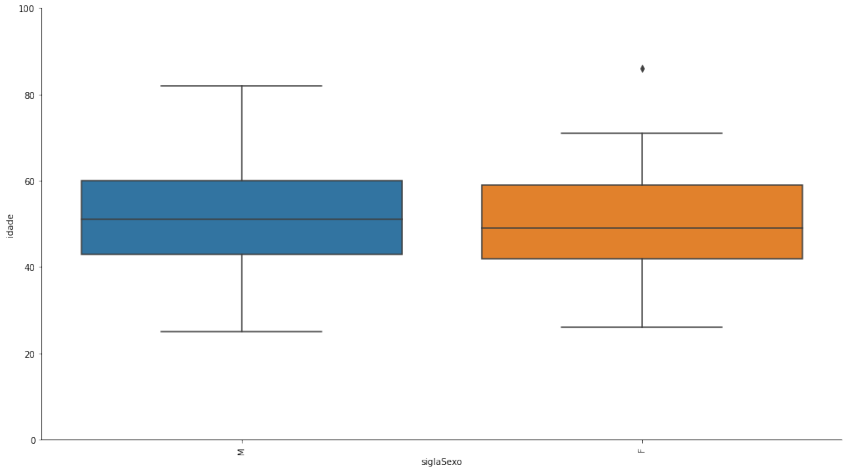


Figure 3.5: Age Distribution Boxplot for Deputies in the Brazilian Chamber of Deputies

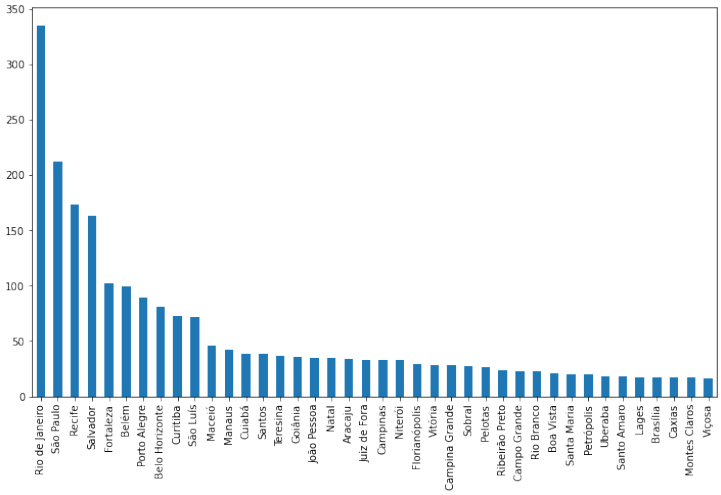


Figure 3.6: Distribution of Deputies by State in the Brazilian Chamber of Deputies

from a minimum of 25 to a maximum of 86 years old Figure 3.6.

In the context of the United States of America, two datasets were utilized for the analysis of voting patterns in 2022. The first dataset provided comprehensive information about elected deputies, while the second dataset summarized the outcomes of the votes cast. We conducted an

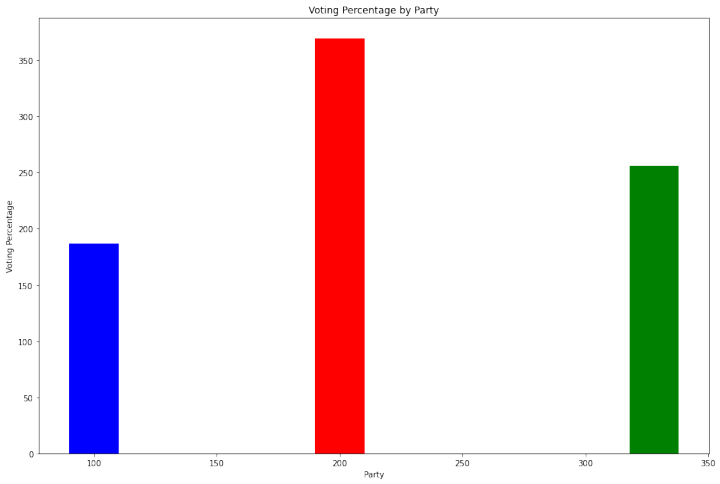


Figure 3.7: Visualization depicting the number of votes cast by party affiliation in the United States.

analysis of the voting percentages by party, which provided intriguing insights into the distribution of votes in the dataset. The party codes served as identifiers, with 100 representing the Democratic Party, 200 representing the Republican Party, and 328 representing the Independent (Figure 3.7).

The Republican Party (code 200) emerged with the highest voting percentage, as depicted by the towering red bar in the plot. This indicates their significant voting strength and larger share of votes compared to the other parties. Following closely behind is the Democratic Party (code 100) with the second-highest voting percentage. Although their voting percentage is lower than that of the Republican Party, the prominent blue bar signifies their substantial voting influence. In contrast, the Independent party (code 328) exhibits the lowest voting percentage, as represented by the relatively shorter green bar.

To further investigate the dynamics of voting, we examined the changes in vote types. In accordance with the Voteview database [14], vote types are classified using codes ranging from 0 to 9. Each code represents a distinct category, as follows:

- 0: Not a member of the chamber when this vote was taken.
- 1: Yea (in favor of the proposal).
- 2: Paired Yea (paired with another member who voted in favor).
- 3: Announced Yea (publicly declared a vote in favor).
- 4: Announced Nay (publicly declared a vote against).

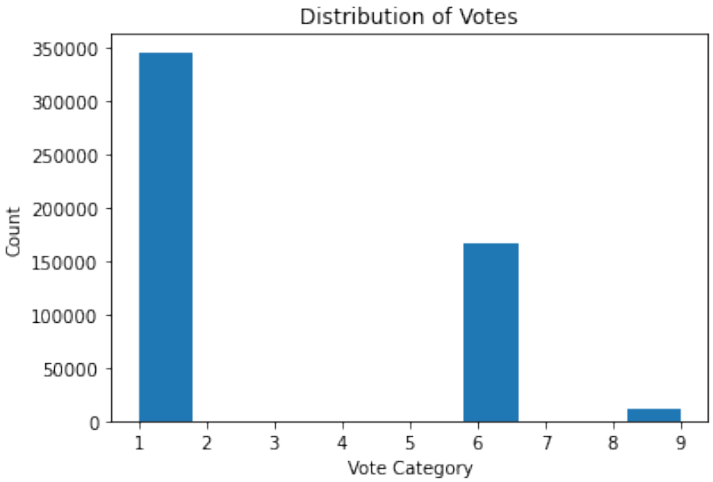


Figure 3.8: Visual representation illustrating the distribution of vote types in the United States.

- 5: Paired Nay (paired with another member who voted against).
- 6: Nay (against the proposal).
- 7 and 8: Present (some Congresses).
- 9: Not Voting (Abstention).

Building upon this analysis, similar to the approach taken with the Brazilian dataset, we sought to understand the distribution of vote types within the United States dataset, as can be seen in Figure 3.8. Upon examining the data, we arrived at a conclusion that aligns with the findings from the Brazilian dataset. The majority of the votes cast in the United States dataset were categorized as "Yea" or "Nay," indicating a clear expression of support or opposition to the proposals. These two vote types dominated the voting outcomes. Additionally, we observed a relatively small percentage of votes categorized as "Not Voting" or "Abstention."

We also conducted an analysis to examine how the Democrats and Republicans are distributed across different states as seen in Figure 3.9. Our findings revealed a distinct concentration of Democratic support in the state of California (CA), while Republican support was predominantly observed in the state of Texas (TX). This analysis highlights the regional variations in political affiliations, with Democrats having a strong presence in California and Republicans being more prevalent in Texas.

In conclusion, the data analysis conducted for both Brazil and the United States provided valuable insights into the voting patterns and dynamics within their respective legislative bodies.

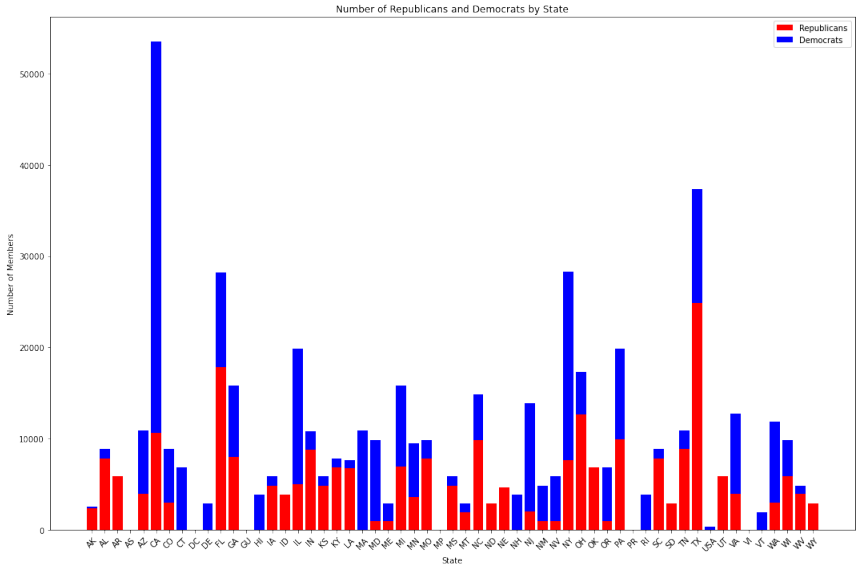


Figure 3.9: Visual representation illustrating the number of votes attributed to the Republican and Democratic parties across different states in the United States.

3.4 Evaluation and Validation of Recommendation Systems

This section comprehensively explores the evaluation and validation process for recommendation systems. It outlines the methodology employed to assess the performance and effectiveness of the developed recommendation system.

3.4.1 Development environment and tools used

Python and Jupyter Notebook formed the core of our thesis project's development environment. Python's versatility, extensive libraries, and ease of use made it the ideal programming language for data manipulation, analysis, and model implementation. Jupyter Notebook's interactive and collaborative nature facilitated our research workflow, allowing us to blend code, visualizations, and explanatory text seamlessly.

The development of the thesis's recommendation algorithm was greatly assisted by the Surprise Library. Surprise offered a complete set of algorithms, assessment measures, and utility functions as a Python library created exclusively for developing and analyzing recommender systems.

To facilitate accessibility, we have shared all of the source code for our thesis project on GitHub [17]. The recommendation algorithm code implementations, data manipulation scripts, and any additional tools or resources used in the research are all available.

3.4.2 Implementation of the recommendation system

In the subject of recommender systems, numerous algorithms have been developed to improve user experience and provide personalized recommendations. Various recommendation algorithms were investigated during the construction of this thesis in order to improve the effectiveness of the recommendation system.

Prominent algorithms in the topic are presented in this section, along with some of their key features.

Algorithm	**Approach**	**Feedback Type**	**Computational Complexity**
KNNBaseline	Collaborative	Explicit & Implicit	High
KNNWithMeans	Collaborative	Explicit & Implicit	Moderate
SVD	Model-Based	Explicit	High
NMF	Model-Based	Explicit	Moderate
CoClustering	Model-Based	Explicit & Implicit	Moderate
SlopeOne	Collaborative	Explicit	Low
BaselineOnly	Collaborative	Explicit	Low
NormalPredictor	Content-Based	N/A	Low

Table 3.1: Comparison of Collaborative Filtering and Content-Based Recommendation Algorithms

Memory-based collaborative filtering algorithms have been examined, including KNNBaseline, KNNWithMeans, SlopeOne, and BaselineOnly. To improve recommendation accuracy, these algorithms account for user and item biases. KNNBaseline includes a baseline estimate, whereas KNNWithMeans corrects for biases using mean ratings. SlopeOne considers differences in item evaluations, whereas BaselineOnly is only based on user and item biases. The capacity of these algorithms to handle both explicit and implicit feedback is advantageous for binary rating suggestions. Additionally, model-based approaches such as SVD, NMF, and CoClustering have been explored as well. SVD is used to divide the user-item matrix into lower-dimensional matrices, which capture latent factors that contribute to ratings. The matrix is factored into non-negative matrices via NMF, making it computationally efficient for large datasets. CoClustering connects users and products based on their similarities. These methods are suitable for handling explicit feedback in binary rating recommendation systems. The NormalPredictor technique was also utilized as a baseline for evaluation. While it predicts scores at random based on the distribution of the training data, it can be used to compare the performance of other methods.

Within the overall setting of this dissertation, we used a precisely established function to test the performance of a varied array of algorithms on the two separate datasets from Brazil and the United States. The main goal was to get profound insights into the delicate interaction between both these datasets and the algorithms under consideration.

The function itself adheres to a well-established and widely accepted paradigm known as the train/test split, a procedure paramount in evaluating the efficacy of recommendation systems.

By virtue of this approach, the algorithms are exposed to a segregated training set, which serves as a conduit for them to assimilate the underlying patterns and intricacies inherent within the data. Consequently, armed with this acquired knowledge, the algorithms undergo a rigorous evaluation on the test set, wherein their predictive capabilities are put to the test.

During the training phase, the algorithms engage in an intricate process of adaptation and refinement, meticulously fine-tuning their internal parameters and architectures. This delicate dance enables the algorithms to effectively encode the latent information within the training set, yielding a learned model capable of capturing the subtle nuances and underlying relationships present within the dataset.

Subsequently, the algorithms are subjected to a comprehensive evaluation process during the testing phase. Employing the knowledge acquired from the training phase, the algorithms are tasked with generating predictions for the test set. These predictions serve as a reflection of the algorithms' ability to discern the underlying binary ratings associated with the dataset.

A pivotal aspect of the function lies within the evaluation metrics employed to quantify the performance of the algorithms. Accuracy, precision, recall, and F1 score augment the evaluation, collectively shedding light on the algorithms' ability to discriminate between positive and negative outcomes.

To facilitate a more granular understanding of the algorithms' performance, the function employs a meticulous conversion scheme, whereby the continuous predicted ratings are transformed into binary values. Employing a threshold of 0.5 as a demarcation point, ratings surpassing this threshold are classified as positive outcomes, while those falling below are regarded as negative outcomes. This binary conversion further enhances the interpretability and utility of the results.

To provide a comprehensive representation of the algorithms' performance, the function produces a meticulously crafted confusion matrix. This matrix intricately outlines the quantities of true positives, false positives, false negatives, and true negatives, acting as a guide to deeply comprehend the algorithms' intricate predictive capacities.

The results derived from the evaluation process are not merely ephemeral entities; rather, they are diligently stored within a structured dictionary. This repository encapsulates the evaluation metrics, each algorithm corresponding to a distinct key within the dictionary. Consequently, this structured representation facilitates efficient analysis and comparison of the algorithms' performance, offering valuable insights for subsequent decision-making.

In pursuit of a cohesive narrative, the function culminates in the transformation of the aforementioned dictionary into a refined data frame. This data frame undergoes meticulous sorting based on the F1 score, an encompassing metric that balances precision and recall. The resulting ordered representation unveils the algorithms' comparative performance, enabling the identification of the most proficient recommendation algorithms within the context of the Brazilian and US datasets.

Harnessing the prowess of this bespoke function, the thesis achieves an unprecedented level of understanding concerning the performance of various algorithms on the Brazilian and US datasets. The assimilated evaluation metrics and the comprehensive confusion matrix collectively contribute to an insightful exploration of the algorithms' efficacy in predicting binary ratings within the context of these datasets.

Chapter 4

Experiments and tests

4.1 Evaluation metrics and methodology

In this study, precision, accuracy, recall, and F1-score were adopted as evaluation metrics to gauge the system's performance for binary rating recommendations. The selection of these metrics was based on their relevance to the problem domain and their capacity to offer comprehensive insights into the system's effectiveness.

Precision (P), is defined as the ratio of true positive predictions (TP) to the total number of positive predictions made by the system (TP + FP).

$$P = \frac{TP}{TP + FP} \tag{4.1}$$

For binary rating recommendation systems, precision indicates the system's accuracy in correctly identifying positive recommendations. A higher precision value signifies greater accuracy and reliability in the system's recommendations, thereby minimizing the occurrence of false positives. Precision is particularly valuable in domains where erroneous recommendations, such as political voting decisions, bear substantial consequences.

Accuracy is defined as the ratio of the number of correct predictions (TP + TN) to the total number of predictions made by the system (TP + TN + FP + FN).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

It represents the overall correctness of the system's recommendations. In the context of binary rating recommendations, accuracy reflects the system's ability to correctly classify both positive and negative recommendations. However, accuracy alone may not provide a comprehensive assessment, particularly when dealing with imbalanced datasets where the number of positive and negative recommendations significantly differs.

Recall (R), also known as sensitivity or true positive rate, is defined as the ratio of true positive predictions (TP) to the total number of actual positive instances (TP + FN).

$$R = \frac{TP}{TP + FN} \tag{4.3}$$

In the context of binary rating recommendation systems, recall measures the system's ability to identify all positive instances correctly. A higher recall value indicates that the system can effectively capture a larger proportion of positive recommendations, minimizing the occurrence of false negatives. Recall is particularly valuable in domains where missing positive recommendations can have significant consequences, such as in medical diagnosis or fraud detection.

F1-score is a harmonic mean of precision and recall, providing a balanced measure of the system's performance by considering both metrics simultaneously. It combines the precision and recall values into a single score, which can be calculated as follows:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4.4}$$

The F1 score considers both precision and recall, making it a suitable metric for evaluating binary rating recommendation systems. It provides a balanced assessment of the system's ability to correctly classify positive recommendations while minimizing false positives and false negatives. A higher F1 score indicates a better trade-off between precision and recall, signifying a more effective recommendation system.

By utilizing precision, accuracy, recall, and F1-score as evaluation metrics, the performance of binary rating recommendation systems can be thoroughly assessed, considering various aspects such as accuracy, reliability, sensitivity, and balance between precision and recall.

4.2 Generation of Realistic Test Data

The development of a realistic setting for the datasets under consideration, which included both Brazilian and American contexts, was of the utmost importance within the framework of this thesis. To that purpose, a meticulously developed function was designed to endow the datasets with a sense of authenticity, closely replicating real-world situations and dynamics.

The overarching objective of this function resided in its ability to imbue the datasets with temporal dimensions, thereby emulating the intricate tapestry of user-item interactions unfolding over time. By meticulously orchestrating the ordering of data points based on their temporal attributes, a faithful representation of the sequential nature of these interactions was achieved. This temporal alignment proved instrumental in facilitating subsequent dataset partitioning and processing.

The multifaceted function featured a twofold methodology. Firstly, the datasets were

meticulously sorted, leveraging the timestamps or any other pertinent temporal indicators at hand. This meticulous ordering of data points in accordance with their temporal occurrence enabled the establishment of a coherent chronology, serving as a testament to the genuine sequence of user-item interactions across time. This preliminary step assumed pivotal significance in the subsequent stages of dataset manipulation and division.

The second facet of the function encompassed the intricate process of segregating the datasets into training and testing subsets, effectively replicating the practical scenario of training recommendation systems by subsequently evaluating their efficacy on previously unseen, more recent data. In an earnest pursuit of simulating a lifelike train-test split, a myriad of strategies were adroitly deployed. Notably, a judicious allocation of a certain percentage of the earliest data for training purposes, while carefully preserving the most recent interactions for testing, was undertaken.

Moreover, the function astutely considered the inherent dynamism within user-item interactions. Employing advanced methodologies, multiple train-test divisions were orchestrated, encapsulating discrete temporal epochs. This ingenious mechanism facilitated a comprehensive evaluation of the recommendation system's performance across diverse stages of its evolutionary trajectory, adeptly adapting to the ever-shifting landscape of user preferences.

By meticulously applying this bespoke function, the datasets emanating from both Brazil and the USA were transmuted into a realm of heightened realism, seamlessly aligning them with the temporal dynamics inherently intertwined with real-world recommendation scenarios. As a result, the thesis was afforded a unique vantage point to meticulously explore and evaluate the intricate algorithms under scrutiny within a context that authentically mirrored the intricate challenges and complexities confronted by recommendation systems in practical, real-world settings.

In order to simulate a hypothetical scenario and assess the sensitivity of the recommendation models, a deliberate choice was made to allocate 80 percent of the datasets as the training set, while reserving the remaining portion for testing. However, in an effort to introduce an element of unpredictability and further challenge the models, a fraction of the test set, represented by a variable l, was included in the training set.

This deliberate inclusion of a percentage of the test set in the training set aimed to simulate a scenario where the recommendation system had access to some recent interactions during the training phase, thereby testing the system's adaptability to new user preferences and evolving item dynamics. By introducing this dynamic element, the models were required to learn and generalize from a combination of historical and relatively recent user-item interactions, effectively assessing their sensitivity to changing patterns and evolving trends.

The value of l was varied between 10%, 5%, and 2% to simulate different levels of sensitivity in the recommendation models. This approach provided valuable insights into the models' ability to adapt and generalize to unseen data, contributing to a more robust evaluation of their performance in realistic settings.

By incorporating these strategic modifications within the meticulously designed function, the datasets underwent a transformative process that not only replicated real-world temporal dynamics but also simulated hypothetical scenarios with varying levels of recent data inclusion. This comprehensive approach, with varying values of l, afforded the thesis a unique opportunity to delve into the intricate algorithms under scrutiny, thoroughly exploring their capabilities within a context that closely mirrored the challenges and complexities confronted by recommendation systems in practical, real-world scenarios. These simulations provided valuable insights into the models' adaptability, generalization, and performance across a range of sensitivity settings, contributing to a more robust evaluation of their efficacy.

Chapter 5

Results

5.1 Evaluation on Available Data

The two tables (Table 5.1 and Table 5.2) illustrate the evaluation metrics for several recommendation algorithms on datasets from Brazil and the United States.

Model	Accuracy	Precision	Recall	F1 Score
BaselineOnly	0.812	0.831	0.833	0.832
CoClustering	0.919	0.929	0.925	0.927
KNNBaseline	**0.924**	**0.936**	**0.928**	**0.932**
KNNWithMeans	0.924	0.936	0.928	0.932
NMF	0.922	0.940	0.920	0.930
NormalPredictor	0.507	0.562	0.549	0.556
SlopeOne	0.812	0.832	0.832	0.832
SVD	0.924	0.935	0.928	0.932

Table 5.1: Evaluation Metrics for Available Data in Brazil

Comparing the two tables, it is interesting to observe the variations in model performance between the Brazilian dataset and the USA dataset. Overall, the models achieved higher performance on the USA dataset compared to the Brazilian dataset across multiple metrics, including accuracy, precision, recall, and F1 score.

The KNNWithMeans and KNNBaseline models consistently performed well in both datasets, showcasing high values for accuracy, precision, recall, and F1 score. These models demonstrate their effectiveness in generating accurate recommendations, regardless of the regional context of the dataset.

It is worth mentioning that several models exhibited significant differences in performance between the two datasets. For instance, the SVD model achieved slightly better accuracy, precision, recall, and F1 score on the Brazilian dataset, while the NMF model outperformed the

Model	Accuracy	Precision	Recall	F1 Score
BaselineOnly	0.798	0.827	0.887	0.856
CoClustering	0.909	0.919	0.949	0.934
KNNBaseline	**0.960**	**0.965**	**0.976**	**0.970**
KNNWithMeans	**0.960**	**0.965**	**0.976**	**0.970**
NMF	0.949	0.965	0.959	0.962
NormalPredictor	0.552	0.677	0.646	0.661
SlopeOne	0.799	0.829	0.886	0.856
SVD	0.959	0.963	0.976	**0.970**

Table 5.2: Evaluation Metrics for Available Data in the USA

USA dataset. These differences suggest that certain models may be more suitable or sensitive to the properties and patterns found in specific regional datasets.

5.2 Evaluation on Realistic Data

Table 5.3 and Table 5.4 present the evaluation metrics for various recommendation algorithms on datasets obtained from Brazil and the United States, respectively. These metrics were calculated using a realistic train-test split, as discussed in the preceding chapter. The results presented here represent the evaluation with the variable n set to 10%.

Model	Accuracy	Precision	Recall	F1 Score
BaselineOnly	0.815	0.814	0.883	0.847
CoClustering	0.941	0.944	0.954	0.949
KNNBaseline	0.816	0.814	0.884	0.848
KNNWithMeans	0.816	0.814	0.884	0.848
NMF	0.942	0.952	0.948	0.950
NormalPredictor	0.508	0.579	0.554	0.566
SlopeOne	0.815	0.814	0.883	0.847
SVD	**0.889**	**0.888**	**0.926**	**0.906**

Table 5.3: Evaluation Metrics for Brazil's Realistic Data Analysis

The evaluation metrics for the Brazil dataset, as shown in Table 5.3, indicate the performance of various recommendation models. The SVD model achieved an accuracy of 0.889, with precision, recall, and F1 score values ranging from 0.888 to 0.926. The SlopeOne model also performed reasonably well with an accuracy of 0.815 and an F1 score of 0.847. NMF exhibited a high accuracy of 0.942 and achieved a balanced performance with precision, recall, and F1 score values around 0.950. However, the NormalPredictor model had a relatively lower accuracy of 0.508, indicating less reliable predictions. Overall, the evaluation suggests that the SVD and NMF models performed well on the Brazil dataset, while the SlopeOne model demonstrated

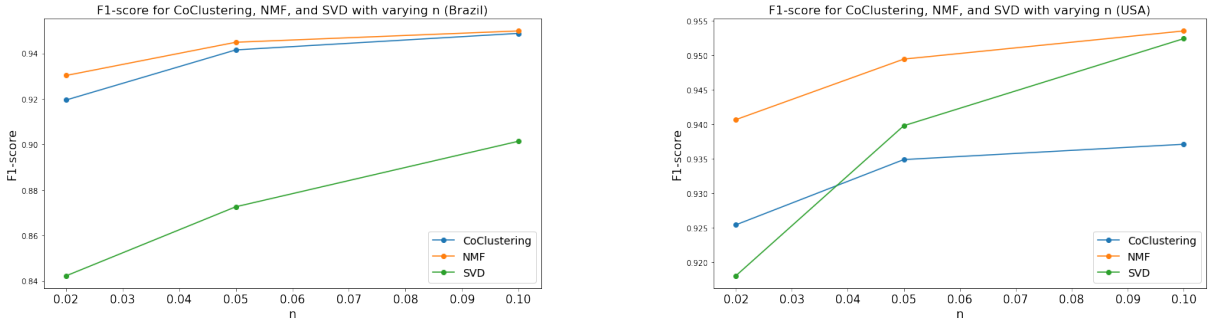
satisfactory performance.

Model	Accuracy	Precision	Recall	F1 Score
BaselineOnly	0.869	0.883	0.951	0.916
CoClustering	0.904	0.918	0.957	0.937
KNNBaseline	0.890	0.906	0.951	0.928
KNNWithMeans	0.890	0.906	0.951	0.928
NMF	**0.932**	**0.953**	**0.956**	**0.954**
NormalPredictor	0.570	0.746	0.642	0.690
SlopeOne	0.869	0.884	0.949	0.915
SVD	0.923	0.930	0.969	0.949

Table 5.4: Evaluation Metrics for USA's Realistic Data Analysis

For the USA dataset, as presented in Table 5.4, the SVD model showcased strong performance with an accuracy of 0.923 and a high recall value of 0.969. The SlopeOne model achieved an accuracy of 0.869 and balanced precision, recall, and F1 score values around 0.915. NMF demonstrated excellent performance with an accuracy of 0.932 and precision, recall, and F1 score values around 0.954. The NormalPredictor model had a relatively lower accuracy of 0.570, indicating its limited predictive capabilities. In summary, the SVD, SlopeOne, and NMF models performed well on the USA dataset, with SVD being particularly notable for its high recall value.

As previously stated, we assessed the performance of many collaborative filtering algorithms for recommendation systems using various n values. The F1-score, a widely used statistic for assessing the algorithm's accuracy and balance of precision and recall, was utilized to analyze the performance. The evaluation employed n values of 0.02, 0.05, and 0.10. In Figure 5.1a and Figure 5.1b we can see the three best-performing algorithms with the n variation.



(a) Variation of F1-score in Brazil

(b) Variation of F1-score in the USA

Figure 5.1: Visual representation illustrating the variation of F1-score when n changes.

The evaluation of algorithms for recommendation systems in the Brazilian and USA datasets revealed consistently high F1 scores for the NMF, SVD, and CoClustering algorithms across different n values. These algorithms demonstrate their effectiveness in providing accurate recommendations. Conversely, the NormalPredictor algorithm consistently obtained lower F1 scores, indicating its limitations in generating accurate recommendations. The NMF, SVD,

and CoClustering algorithms can be considered strong choices due to their consistently high performance as can be seen in both Figure 5.1a and Figure 5.1b.

In summary, the evaluation of realistic data confirmed the effectiveness of the SVD and NMF models for the Brazil dataset, while the SlopeOne and NMF models performed well on the USA dataset. These results highlight the importance of considering regional differences and dataset characteristics when selecting recommendation algorithms.

5.3 Discussion

The evaluation of available data demonstrated variations in the performance of the recommendation algorithms between the Brazilian and USA datasets. The USA dataset consistently yielded higher performance metrics compared to the Brazilian dataset. This difference can be attributed to several factors, including variations in data distribution, data sparsity, and data quality. Being more representative and diverse, the USA dataset allowed the models to capture meaningful patterns and generate accurate recommendations. In contrast, the Brazilian dataset, which may have been sparser, posed challenges for the models, leading to slightly lower performance metrics

However, the evaluation on realistic data painted a different picture. The performance metrics for the Brazilian dataset improved significantly, and some recommendation algorithms even outperformed their counterparts on the USA dataset. This result can be attributed to the realistic data's better representation of user preferences and behaviors, as well as its higher data quality. The realistic dataset, obtained through a careful train-test split that mimicked real-world scenarios, provided a more accurate evaluation of the models' performance. The algorithms were able to leverage the denser and more relevant information in the realistic dataset, resulting in improved accuracy and precision.

Chapter 6

Conclusions and Future Work

In conclusion, politics has a huge impact on cultures and shapes their opinions on critical topics. Making educated and thoughtful decisions during political events is critical for the advancement of the democratic process in democratic countries such as Brazil and the United States. However, both countries' political systems confront issues, such as polarisation, a diminished central state, and a lack of faith in the government.

We did a detailed literature analysis and investigated case studies to investigate the potential of recommendation systems in politics. The review highlighted the use of big data and recommendation algorithms in political campaigns to gather public opinion and influence voting behavior, as proven by cases from the United States, Brazil, and India.

Based on this understanding, our research intends to analyze deputies' behavior and establish a recommended approach for measuring voting patterns and trends of deputies and political parties in Brazil and the United States. We hope to obtain insights into the political dynamics and decision-making processes within legislative bodies by utilizing suggestion methodologies.

The significance of our research lies in the potential to bridge the gap between recommendation systems and the political domain, thereby empowering individuals and fostering informed decision-making processes. Through the application of data-driven methodologies, we aspire to provide actionable insights that can enrich political discourse, strengthen democratic practices, and ultimately contribute to the advancement of societies in Brazil and the United States.

As our thesis concludes, we acknowledge that this is the beginning of a multifaceted journey. Further research and exploration in this field hold the promise of uncovering new dimensions of recommendation systems in politics, opening doors to enhanced public engagement, refined policy-making, and ultimately, the promotion of thriving democratic societies.

6.1 Future Work

The developed recommendation system for political decision-making has demonstrated its potential to assist individuals with informed choices. However, there are several areas that warrant further exploration and development. The following are potential avenues for future work:

- **Platform Accessibility:** Broaden the reach and impact of the recommendation system by making it accessible through widely available platforms such as public websites or mobile applications, thereby enabling a larger audience to engage in informed political decision-making.
- **Diverse Political Contexts:** Extend the scope of the investigation to encompass countries with diverse political landscapes. By tailoring the recommendation system to cater to varying political ideologies, more accurate and relevant recommendations can be provided to users.
- **User Feedback and Iterative Enhancement:** Incorporate mechanisms for user feedback, such as surveys or user ratings, to gain insights into the quality and relevance of recommendations. Utilize this feedback to iteratively refine and enhance the recommendation algorithms.
- **Evaluation and Validation:** Undertake rigorous evaluation and validation studies, including user-centric assessments, surveys, and comparative analyses of user behaviors and outcomes. These efforts will ascertain the effectiveness, impact, and contribution of the system to informed decision-making and political engagement.
- **Temporal Variability in Recommendation Systems:** Explore the applicability of recommendation systems in contexts where features exhibit temporal variations, particularly in the case of Content-Based Recommendation Systems. Investigate how these systems can adapt to changing dynamics over time.

As we transition from the culmination of our current research to the horizon of future possibilities, these areas beckon for further investigation and innovation, promising to illuminate new avenues for the synergy of recommendation systems and the political domain.

References

[1] Brent Smith and Greg Linden. Two decades of recommender systems at amazon.com. *IEEE Internet Computing*, 21(3):12–18, 2017. doi:10.1109/MIC.2017.72.

[2] Dr. Christopher Sabatine and Jon Wallace. Democracy in brazil - chatham house – international affairs think tank, Jan 2023.

[3] Anna Kaufman. What are the three branches of government? executive, judicial, legislative wings explained., Oct 2022.

[4] Carole Cadwalladr. Facebook suspends data firm hired by vote leave over alleged cambridge analytica ties, 2018.

[5] Matthew L. Layton, Amy Erica Smith, Mason W. Moseley, and Mollie J. Cohen. Demographic polarization and the rise of the far right: Brazil's 2018 presidential election. *Research & Politics*, 2021. doi:10.1177/2053168021990204.

[6] Michael Safi. India's 'big data' election: 45,000 calls a day as pollsters target age, caste and religion, 2017.

[7] Ian MacKenzie, Chris Meyer, and Steve Noble. How retailers can keep up with consumers, Oct 2013.

[8] Charu C. Aggarwal. *Recommender Systems The Textbook*. Springer International Publishing, march 2016. ISBN: 978-3-319-29657-9.

[9] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems, 2004. ISSN: 1046-8188. doi:10.1145/963770.963772.

[10] Aakanksha NS. Recommender systems: Matrix factorization using pytorch, Nov 2020.

[11] Nick Anstead. Data-driven campaigning in the 2015 uk general election, Jul 2017.

[12] Eric Hellweg. 2012: The first big data election, Aug 2014.

[13] Dados Abertos da Câmara dos Deputados. [link].

[14] Voteview. [link].

[15] Craig Stedman, Ed Burns, and Mary K. Pratt. What is data preparation? an in-depth guide to data prep, Feb 2022.

[16] Rodrigo Nunes Moni da Silva, Andre Spritzer, and Carla Dal Sasso Freitas. Visualization of roll call data for supporting analyses of political profiles. pages 150–157, 2018. doi:10.1109/SIBGRAPI.2018.00026.

[17] Github Repository. [link].

[18] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Recommender Systems Handbook*, pages 1–35. 10 2010. ISBN: 978-0-387-85819-7. doi:10.1007/978-0-387-85820-31-1.

[19] Baptiste Rocca. Introduction to recommender systems, Jun 2019.

[20] Maruti Techlabs. Types of recommendation systems & their use cases, Aug 2021.

[21] Victor S Bursztyn, Marcelo G Nunes, and Daniel R Figueiredo. How Brazilian congressmen connect: homophily and cohesion in voting and donation networks. *Journal of Complex Networks*, 8(1), 02 2020. ISSN: 2051-1329. doi:10.1093/comnet/cnaa006.

[22] Breno de Matos, Carlos H. Ferreira, and Jussara M. Almeida. Analisando a governabilidade presidencial a partir de padrões de homofilia na câmara dos deputados: Estudos de casos no brasil e nos eua. *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2018. doi:10.5753/brasnam.2018.3592.

[23] Carlos Ferreira, Fabricio Murai, Breno Matos, and Jussara Almeida. Modeling dynamic ideological behavior in political networks. pages 1–14, 08 2019.

[24] Sanket Doshi. Brief on recommender systems, Feb 2019.

[25] Cem Dilmegani. Recommendation systems: Applications and examples in 2023.

[26] Salma Ghoneim. Accuracy, recall, precision, f-score & specificity, which to optimize on?, Apr 2019.

[27] Sahil Mankad. A tour of evaluation metrics for machine learning, Dec 2020.

[28] Bin Wang. Comparison of user-based and item-based collaborative filtering, May 2022.

[29] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019. ISSN: 1877-0509. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. doi:https://doi.org/10.1016/j.procs.2019.11.177.

[30] Robin Burke. Hybrid recommender systems: Survey and experiments - user modeling and user-adapted interaction, Nov 2002.