Captura y análisis de datos para el estudio de la evolución de las tendencias políticas en las elecciones de EE.UU.

Web scraping and data analysis for the study of the evolution of political trends in US elections

Trabajo de Fin de Grado
Curso 2024–2025

Autor
Laura Rodrigo Cañete

Director
Rafael Caballero Roldán

Calificación: 10 (SB)

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

Captura y análisis de datos para el estudio de la evolución de las tendencias políticas en las elecciones de EE.UU.

Web scraping and data analysis for the study of the evolution of political trends in US elections

Trabajo de Fin de Grado en Ingeniería Informática

Autor
Laura Rodrigo Cañete

Director
Rafael Caballero Roldán

Convocatoria: *Junio 2025*

Calificación: 10 (SB)

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

20 de Mayo de 2025

Dedication

To my parents, who will be the people who read this with the most enthusiasm.

Acknowledgments

I want to thank my tutor, Rafael Caballero Roldán, for being an excellent teacher and helping me in all that he could. He has trusted my decisions and encouraged me while respecting my organization and timings. He has brought laughter to every meeting and made this process into something I actually enjoyed and learned from.

Abstract

Web scraping and data analysis for the study of the evolution of political trends in US elections

This study investigates political polarization in the United States by analyzing user behavior on X (formerly Twitter) during the 2016, 2020, and 2024 presidential elections. We developed a custom web scraping tool to collect real-time data related to the 2024 election, complementing previously acquired datasets from 2016 and 2020. To infer political alignment, we encoded each user's tweet history into vector representations using sentence-transformer models, then constructed similarity-based graphs where users are nodes connected by opinion proximity and shared content. We manually labeled a small set of users and propagated these labels through the network using various contagion strategies. The method was evaluated against alternative classification approaches and shown to be both accurate and robust. We applied it to analyze discourse strategies and voter behavior across the three election cycles, uncovering significant shifts in attention focus, content format, and partisan engagement. This graph-based method enables large-scale classification of users as Democrat or Republican and improves the understanding of polarization through social media data.

Keywords

Twitter, political science, political polarization, data analytics, web scraping, inference, vector embedding, graph, label propagation

Resumen

Captura y análisis de datos para el estudio de la evolución de las tendencias políticas en las elecciones de EE.UU.

Este estudio analiza la polarización política en Estados Unidos mediante el estudio del comportamiento de los usuarios en X (anteriormente Twitter) durante las elecciones presidenciales de 2016, 2020 y 2024. Se desarrolló una herramienta personalizada de *web scraping* para recopilar datos en tiempo real relacionados con las elecciones de 2024, complementando así conjuntos de datos previamente obtenidos para 2016 y 2020. Para inferir la alineación política, se codificó el historial de *tweets* de cada usuario en representaciones vectoriales utilizando modelos *sentence-transformer*, y se construyeron grafos basados en similitud, donde los nodos representan usuarios conectados por proximidad de opinión y contenido compartido. Se etiquetó manualmente a un pequeño conjunto de usuarios y se propagaron dichas etiquetas por la red mediante distintas estrategias de contagio. El método fue comparado con otros enfoques de clasificación y demostró ser preciso. Lo aplicamos para analizar las estrategias de discurso y el comportamiento de los votantes a lo largo de los tres ciclos electorales, detectando cambios importantes en el enfoque de atención, el formato del contenido y la participación según el partido político. Este enfoque permite clasificar a gran escala a los usuarios como demócratas o republicanos y mejora la comprensión de la polarización a través de los datos de redes sociales.

Palabras clave

Twitter, ciencia política, polarización política, análisis de datos, web scraping, inferencia, vector embedding, grafo, propagación de etiquetas

Table of Contents

Chapter 1

Introduction

Chapter Summary: This introductory chapter establishes the context and significance of studying political polarization, particularly through the lens of social media platforms like X during U.S. presidential elections. It outlines the primary objectives of the thesis, which include collecting extensive user data, developing a method for classifying users by their political affiliation, and conducting a longitudinal analysis of political trends across multiple election cycles. The chapter details the structured work plan undertaken to achieve these goals, addressing challenges encountered during data acquisition and outlining the chosen analytical approach. Finally, it displays the upcoming thesis structure and lists the key technologies that facilitated the project's development.

1.1. Motivation

Political polarization is a major issue in today's democracies, especially in the U.S., where strong divisions between political groups affect how people vote and talk about politics. Social media platforms, particularly X (formerly Twitter), have played a central role in shaping public opinion during the 2016, 2020, and 2024 presidential elections. The platform has served as a key space for political expression, amplified by figures like Donald Trump and, more recently, Elon Musk –whose influence on the platform's moderation policies and public discourse has drawn significant attention–. As access to X's official API has become increasingly restricted, researchers have turned to web scraping to continue extracting meaningful data at scale.

In this context, modeling polarization with concrete, user-level data offers a powerful mean to study how opinions form, cluster, and evolve. Central to this effort is the ability to classify users by political affiliation. In view of this, this project examined X to analyze and detect a user's political stance. This study collected datasets and labeled a user's stance using a graph-based label propagation approach to determine whether a user was a Democrat or a Republican during key electoral moments.

These studies are important because learning about users' political positions and their activity on social media helps shape campaign strategies and can influence election outcomes as many authors like Dimitrova and Matthes (2018) or Marret (2020) point out.

1.2. Objectives

The main goal of this study is to analyze political opinion in the United States by using user-generated data from the social network X. To achieve this, we focus on three key objectives:

- **Data collection:** Collect tweets related to the 2024 U.S. presidential election using web scraping. These tweets complement those previously obtained from the same users during the 2016 and 2020 elections.
- **User classification:** Develop a method to classify users as supporters of either the Democrat or the Republican Party based on their tweet history. This classification is based on the aggregation and analysis of each user's tweet content, using vector embedding and graph-based techniques.
- **Longitudinal analysis:** Study the evolution of political trends across the three election years, including shifts in user alignment and changes in engagement levels with different candidates over time.

1.3. Work Plan

In order to fulfill these objectives, we followed a structured work plan.

The first step was to obtain the necessary data. Datasets from the 2016 and 2020 elections were provided by a research team, originally collected using Twitter's API. For the 2024 election, we initially planned to use the same method to collect information from the users that had participated in the last two elections but, due to recent changes in X's access policy, this was no longer possible. We explored alternative APIs, but all were either prohibitively expensive or did not offer the flexibility we needed. We also reviewed several open-source scraping projects on GitHub, but found the code often poorly documented or unreliable (because of the fragility and change sensitivity typical of web scraping codes). As a result, we decided to implement our own web scraping solution, believing it would also add value to the project. This phase was particularly time-consuming and involved extensive trial and error, but the scraper was successfully completed in early November and launched one week after the 2024 election. The objective was to collect data covering the interval from the week preceding to the week following the election day, both inclusive. This scraping task was time-sensitive, as older tweets became harder to retrieve, requiring significant computational resources and parallel scraping across multiple machines.

Once the full dataset was collected and processed, we turned to the analysis phase, where we wanted to determine which political party a user primarily supports.

We researched on the different options and although we initially considered using sentiment analysis, we experimented with various libraries and aggregation methods without obtaining satisfactory results. Therefore, we tried a new approach based on vector embeddings: we merged all tweets from each user into a single text string, encoded them using sentence-transformer models, and built a graph where users were nodes connected based on cosine similarity and shared tweets. A small set of users was manually labeled as Democrat or Republican, and then political alignment was propagated through the graph using various contagion strategies, as supported by the literature. This produced colored graphs for each election year. We then evaluated the results to check the accuracy of the method, testing several small adjustments to improve the process. Here a variant of the method that only used one of the propagation strategies was found to be strikingly more precise and was selected to be applied in the Results Chapter.

Finally, we analyzed the resulting user classifications and other user data to observe how political alignment evolved over time. This included identifying typical user profiles for each party and election year, as well as drawing insights on how campaign strategies may have shifted due to the change in the number of tweets posted. Additionally, we conducted some brief descriptive analysis on how tweet formats and behaviors have changed over the three election cycles.

1.4. Thesis Structure

This thesis is structured into eight main chapters, guiding the reader through the research on political opinion inference from social media data. Chapter 1, "Introduction", sets the stage by outlining the motivation, objectives, and overall work plan of the project, in addition to presenting this structural overview and detailing the development technologies used. Chapter 2, "Introducción" is a translation of the first chapter. Chapter 3, "State of the Art", provides a comprehensive review of existing research, focusing on methods and challenges for data collection from X and computational approaches for inferring political affiliation. Moving into the practical aspects, Chapter 4, "Data Retrieval", explains the process of acquiring the datasets for the 2016, 2020, and 2024 U.S. presidential elections, including insights into the original data and the custom web scraping system developed to overcome recent API restrictions. Chapter 5, "From tweets to political opinions", explains the core methodology, detailing the framework assumptions, how users are represented through vector embeddings, the creation of graph connections, and the label propagation mechanisms employed for classification. This chapter also includes a thorough discussion of methodological choices and validation. Chapter 6, "Results", presents the key findings, categorized by user demographics and political leanings, content focus and messaging strategies, and the evolution of tweet composition. Finally, Chapter 7, "Conclusions and Future Work", summarizes the principal conclusions drawn from this research and proposes various ideas for future investigations. Chapter 8, "Conclusiones y Trabajo Futuro", is the translation of the latter.

1.5. Project development and technologies

Throughout the project, we maintained an organized workflow supported by weekly meetings, collaborative sessions on Google Meet, and a Trello board to assign and track tasks. Code was shared via GitHub, while datasets and working documents were stored and synchronized using OneDrive. All development and analysis were conducted in Python with Jupyter notebooks, integrating tools such as Selenium for scraping (The Selenium Project, 2024), MongoDB Atlas for concurrent access to data (MongoDB, Inc., 2024), Excel for manual classification, Hugging Face (Hugging Face, 2024), Flair (Akbik et al., 2019) and other Python libraries for modeling, and Neo4j (Neo4j, Inc., 2024b) for graph visualization and label propagation through Cypher queries (Neo4j, Inc., 2024a).

All code developed for this project is publicly available on GitHub at: `https:/github.com/LauraRodrigoCanete/US_elections_analysis`.

Chapter 2

Introducción

Resumen: Este capítulo introductorio establece el contexto y la importancia de estudiar la polarización política, particularmente a través de plataformas de redes sociales como X durante las elecciones presidenciales de EE. UU. Describe los objetivos principales del trabajo, que incluyen la recopilación de los datos de usuarios, el desarrollo de un método para clasificar a los usuarios por su afinidad política y la realización de un análisis longitudinal de las tendencias políticas a lo largo de múltiples ciclos electorales. El capítulo detalla el plan de trabajo llevado a cabo para lograr estos objetivos, abordando los desafíos encontrados durante la adquisición de datos y especificando el enfoque analítico elegido. Finalmente, presenta la estructura del documento y enumera las tecnologías clave que facilitaron el desarrollo del proyecto.

2.1. Motivación

La polarización política es un problema importante en las democracias actuales, especialmente en EE. UU., donde las fuertes divisiones entre grupos políticos afectan cómo la gente vota y habla de política. Las plataformas de redes sociales, particularmente X (anteriormente Twitter), han jugado un papel central en la formación de la opinión pública durante las elecciones presidenciales de 2016, 2020 y 2024. La plataforma ha servido como un espacio clave para la expresión política, amplificada por figuras como Donald Trump y, más recientemente, Elon Musk, cuya influencia en las políticas de moderación de la plataforma y el discurso público ha atraído una atención significativa. A medida que el acceso a la API oficial de X se ha vuelto cada vez más restringido, los investigadores han recurrido al *web scraping* para seguir extrayendo datos significativos a gran escala.

En este contexto, modelizar la polarización con datos concretos a nivel de usuario ofrece un medio potente para estudiar cómo se forman, agrupan y evolucionan las opiniones. Fundamental para este esfuerzo es la capacidad de clasificar a los usuarios por afiliación política. En vista de esto, este proyecto examinó X para analizar e identificar la postura política de un usuario. Este estudio recopiló datos y etiquetó la postura de un usuario utilizando un enfoque de propagación de etiquetas basado en

grafos para determinar si un usuario era demócrata o republicano durante momentos electorales clave.

Estos estudios son importantes porque aprender sobre las posiciones políticas de los usuarios y su actividad en las redes sociales ayuda a diseñar estrategias de campaña y puede influir en los resultados electorales, como señalan muchos autores como Dimitrova and Matthes (2018) o Marret (2020).

2.2. Objetivos

El objetivo principal de este estudio es analizar la opinión política en Estados Unidos utilizando datos generados por usuarios de la red social X. Para lograr esto, nos centramos en tres objetivos clave:

- **Recopilación de datos:** Recopilar tuits relacionados con las elecciones presidenciales de EE. UU. de 2024 utilizando *web scraping*. Estos tuits completan a los obtenidos previamente de los mismos usuarios durante las elecciones de 2016 y 2020.
- **Clasificación de usuarios:** Desarrollar un método para clasificar a los usuarios como partidarios del Partido Demócrata o Republicano basándose en su historial de tuits. Esta clasificación se basa en la agregación y el análisis del contenido de los tuits de cada usuario, utilizando transformaciones vectoriales y técnicas basadas en grafos.
- **Análisis longitudinal:** Estudiar la evolución de las tendencias políticas a lo largo de los tres años electorales, incluyendo cambios en la alineación de los usuarios y en los niveles de participación con diferentes candidatos a lo largo del tiempo.

2.3. Plan de Trabajo

Para cumplir con estos objetivos, seguimos un plan de trabajo estructurado.

El primer paso fue obtener los datos necesarios. Las bases de datos de las elecciones de 2016 y 2020 fueron proporcionadas por un equipo de investigación, originalmente recopiladas utilizando la API de Twitter. Para las elecciones de 2024, inicialmente planeamos utilizar el mismo método para recopilar información de los usuarios que habían participado en las dos elecciones anteriores, pero, debido a cambios recientes en la política de acceso de X, esto ya no fue posible. Exploramos APIs alternativas, pero los costes eran demasiado altos o no ofrecían la flexibilidad que necesitábamos. También revisamos varios proyectos de scraping de código abierto en GitHub, pero encontramos que el código a menudo estaba mal documentado o era poco fiable (debido a la fragilidad y sensibilidad al cambio típica de los códigos de *web scraping*). Como resultado, decidimos implementar nuestra propia solución de *web scraping*, creyendo que también añadiría valor al proyecto. Esta fase fue particularmente laboriosa e implicó mucha prueba y error, pero el programa se completó

con éxito a principios de noviembre y se lanzó una semana después de las elecciones de 2024. El objetivo era recopilar datos que cubrieran el intervalo desde la semana anterior hasta la semana posterior al día de las elecciones, ambas inclusive. Esta tarea era urgente, ya que los tuits más antiguos se volvían más difíciles de recuperar, lo que requirió importantes recursos computacionales de múltiples ordenadores.

Una vez que se recopiló y procesó el conjunto de datos completo, pasamos a la fase de análisis, donde queríamos determinar qué partido político apoya principalmente un usuario. Investigamos las diferentes opciones y, aunque inicialmente consideramos usar análisis de sentimientos, experimentamos con varias bibliotecas y métodos de agregación sin obtener resultados satisfactorios. Por lo tanto, probamos un nuevo enfoque basado en transformaciones vectoriales: fusionamos todos los tuits de cada usuario en una única cadena de texto, los codificamos usando modelos de transformadores de texto y construimos un grafo donde los usuarios eran nodos conectados en base a la similitud del coseno y los tuits compartidos. Un pequeño conjunto de usuarios se etiquetó manualmente como demócrata o republicano, y luego la alineación política se propagó a través del grafo utilizando varias estrategias de contagio, como lo respalda la literatura. Esto produjo grafos coloreados para cada año electoral. Luego evaluamos los resultados para verificar la precisión del método, probando pequeños ajustes para mejorar el proceso. Aquí, una variante del método que solo usaba una de las estrategias de propagación resultó ser sorprendentemente más precisa y se seleccionó para aplicarse en el capítulo de resultados.

Finalmente, analizamos las clasificaciones de usuarios resultantes y otros datos de usuario para observar cómo evolucionó la alineación política a lo largo del tiempo. Esto incluyó la identificación de perfiles de usuario típicos para cada partido y año electoral, así como la obtención de información sobre cómo pueden haber cambiado las cantidades de tuits publicados por los usuarios como estrategia de campaña. Además, realizamos un breve análisis descriptivo sobre cómo han cambiado los formatos y comportamientos de los tuits a lo largo de los tres ciclos electorales.

2.4. Estructura de la Tesis

Esta tesis se estructura en ocho capítulos principales, guiando al lector a través de la investigación sobre la inferencia de la opinión política a partir de datos de redes sociales. El Capítulo 1, “Introduction”, sienta las bases exponiendo la motivación, los objetivos y el plan de trabajo general del proyecto, además de presentar la organización estructural y detallar las tecnologías de desarrollo utilizadas. El Capítulo 2, “Introducción”, es una traducción del primer capítulo. El Capítulo 3, “State of the Art”, proporciona una revisión exhaustiva de la investigación existente, centrándose en los métodos y desafíos para la extracción de datos de X y los enfoques computacionales para inferir la afiliación política. Pasando a los aspectos prácticos, el Capítulo 4, “Data Retrieval”, explica el proceso de adquisición de los datos para las elecciones presidenciales de EE. UU. de 2016, 2020 y 2024, incluyendo información sobre los datos originales y el sistema personalizado de *web scraping* desarrollado para superar las recientes restricciones de la API. El Capítulo 5, “From tweets to political opinions”, explica la metodología central, detallando los supuestos del

marco, cómo se representan los usuarios mediante transformaciones vectoriales, la creación de conexiones de grafos y los mecanismos de propagación de etiquetas empleados para la clasificación. Este capítulo también incluye una discusión exhaustiva de las elecciones metodológicas y la validación. El Capítulo 6, "Results", presenta los hallazgos clave, categorizados por la demografía del usuario y las inclinaciones políticas, el enfoque del contenido y las estrategias de mensajería, y la evolución de la composición de los tuits. Finalmente, el Capítulo 7, "Conclusions and Future Work", resume las principales conclusiones extraídas de esta investigación y propone diversas vías para futuras investigaciones. El Capítulo 8, "Conclusiones y Trabajo Futuro", es la traducción de este último.

2.5. Desarrollo del proyecto y tecnologías

A lo largo del proyecto, mantuvimos un flujo de trabajo organizado apoyado por reuniones semanales, sesiones colaborativas en Google Meet y un tablero de Trello para asignar y seguir tareas. El código se compartió a través de GitHub, mientras que las bases de datos y los documentos de trabajo se almacenaron y sincronizaron utilizando OneDrive. Todo el desarrollo y análisis se realizó en Python con Jupyter notebooks, integrando herramientas como Selenium para scraping (The Selenium Project, 2024), MongoDB Atlas para acceso concurrente a datos (MongoDB, Inc., 2024), Excel para clasificación manual, Hugging Face (Hugging Face, 2024), Flair (Akbik et al., 2019) y otras bibliotecas de Python para modelado, y Neo4j (Neo4j, Inc., 2024b) para visualización de grafos y propagación de etiquetas a través de consultas Cypher (Neo4j, Inc., 2024a).

Todo el código desarrollado para este proyecto está disponible públicamente en GitHub en:

`https:/github.com/LauraRodrigoCanete/US_elections_analysis`.

Chapter 3

State of the Art

Chapter Summary: In this chapter we will summarize and outline the existing research on the two main areas of the project: methods and challenges for data collection from X (formerly Twitter), and computational approaches for inferring political affiliation from users' social network data. First, it traces the evolution of data extraction methods from X, highlighting the recent shift from API access to web scraping, and discussing the associated ethical, practical, and legal considerations. Second, the computational approaches reviewed include network-based methods that leverage social homophily, content-based methods that analyze user-generated text, and hybrid approaches that combine both. Finally, it also explores some modern techniques such as contagion algorithms and longitudinal perspectives, demonstrating how current methodologies map online users onto the political spectrum.

3.1. Data Extraction from X

In this section we cover the evolution of access methods to the X social network platform, primarily the shift to web scraping as well as the ethical, practical and legal considerations of this practice.

Twitter has long been a popular data source for social media research due to its historically open data policies. In the early 2010s, Twitter provided researchers easy access to large volumes of data via public APIs. Researchers could tap the Streaming API (which offered a free 1 % sample of tweets) or the Search API, making it straightforward to download tweets at scale. In particular, the Python library *Tweepy* (Roesslein, 2009) became the most popular way of downloading tweets and was used extensively by researchers. This was the method used to retrieve the data of the 2016 and 2020 elections.

However, these API-based collections were not without bias. Studies, like Trezza (2023), argue that the data retrieved through Twitter's public APIs can be non-random and incomplete, leading to concerns about representativeness. They claim that Twitter's documentation acknowledged that the Search API does not index all

tweets, and the free Streaming API caps results when query volume exceeds roughly 1 % of global tweet traffic. Consequently, findings drawn from such API samples were at risk of bias –certain hashtags or user groups might be over or underrepresented–, affecting the integrity of research conclusions.

In the aftermath of high-profile data scandals (e.g. the 2018 Cambridge Analytica incident on Facebook[1]), many platforms tightened data access. Twitter remained relatively open longer than some peers, but in 2023 a major shift occurred. Twitter was re-branded as X and announced the end of free academic API access, raising its prices considerably, with Enterprise API pricing starting at $42,000/Month according to the official website X Corp. (2024).

These changes, including the revocation of existing academic tokens, made X's API cost-prohibitive for most researchers, abruptly stopping many ongoing projects. By mid-2023, people who relied on X data were left with few official channels to collect new data and this prompted a search for alternative methods. The primary response has been a turn to web scraping, i.e. writing scripts or using browser automation to pull data directly from X's web interface. Scraping techniques (e.g. using Python tools like Selenium or requests) allow bypassing the API by extracting tweets and metadata from publicly available HTML or JSON responses. Indeed, observers as Brown et al. (2024) anticipate "an increase in research relying on scraping data from the web" in light of these restrictions.

The move from API use to scraping raises important ethical, legal, and methodological questions. Unlike official APIs –which enforce rate limits and terms of service compliance– web scraping can easily violate platform's terms of service and potentially user privacy expectations. Researchers (Trezza (2023)) have described scraping as a "necessary evil" under new constraints, but emphasize that it must be done under strict data management conditions to protect individuals' privacy. Key concerns include obtaining truly public data, avoiding collection of sensitive personal information, and anonymizing or aggregating data to minimize harm. Additionally, the representativeness of scraped datasets can be an issue. If scraping is done via search queries or specific user timelines, the sample may be skewed towards certain topics or user segments. Experts as Mensah (2023) note that as access to comprehensive data diminishes, transparency about data collection methods is vital to assess any biases introduced.

Regarding the legal aspects involved, scraping data for research operates within a complex and evolving legal landscape, requiring case-by-case analysis. Scraping public data generally carries lower legal risk than accessing private or restricted content, but privacy concerns still demand careful handling of any personal information collected. According to Brown et al. (2024), researchers should limit data collection to what is necessary, focus on publicly available tweet data, ensure transparency and stay informed of changing regulations.

In summary, the current state of X's data extraction is one of adaptation: re-

[1] A former Cambridge Analytica employee disclosed that Facebook had been involved in the misuse of user data. This incident became known as the Facebook–Cambridge Analytica data scandal. On March 17, 2018, the Guardian and New York Times broke the story, saying that the company had utilized 50 million Facebook profiles to do their modeling. The Federal Trade Commission ended up fining Facebook for privacy violations (Hern and Pegg, 2018).

searchers are developing scraping tools and ethical guidelines to continue using X's data, while acknowledging the limitations and biases that come with this post-API era.

3.2. Inference of Political Affiliation from Users

In this section we will review some of the most popular methods for automatic classification of users, such as network-based approaches (involving graphs) and content-based approaches (related to text analysis). We will also discuss hybrid methods that involve the two last approaches and mention other modern methods such as contagion algorithms (methods that use label propagation) and longitudinal perspectives (methods that consider how variables evolve over time). The method that we will apply in this project shares characteristics with all of the previous.

Inferring a user's political affiliation (e.g. Democrat vs. Republican in the U.S. context) from their social media activity has become an important computational social science task. The problem is challenging because political alignment is a latent attribute; researchers must rely on clues from content and network structures to make predictions.

One influential line of work are the network-based approaches that study social network homophily: the tendency of like-minded individuals to connect or engage with each other (Bisgin et al., 2012). In X, users with similar political views often form clusters in the retweet or follow network. Early studies showed that these networks can be used to accurately classify ideology. For instance, Conover et al. (2011) applied a label propagation algorithm to the X retweet graph during the 2010 U.S. midterms, spreading known labels (e.g. accounts of known partisans) through the network. This graph-based approach outperformed classifiers that used tweet text based machine learning methods alone. Similarly, other researchers have used the following network: if a user follows many political figures from one party, one can infer that user's alignment. It is common to find methods that model relationships of users that follow each other, some using dimensionality reduction, like An et al. (2012), or Bayesian models, as Barberá (2015), to place users on an ideological spectrum. These approaches treat the network as a graph where edges indicate potential ideological affinity; then through techniques like label propagation or graph partitioning, unknown users can be assigned to communities (left-leaning or right-leaning). The underlying assumption –supported by empirical evidence– is that political information diffuses in the network: users who retweet each other or follow the same leaders tend to share ideology, allowing a form of contagion-based inference whereby a few seed labels spread to many nodes.

Another major strategy focuses on the text content of users' posts. Political ideology often manifests in the language people use, the topics they discuss, or the hashtags they adopt. Traditional content-based methods used features like keywords (e.g. political hashtags or phrases), linguistic style, or bag-of-words representations of a user's tweets to train classifiers. With advances in NLP (Natural Language Processing), modern approaches use vector-based representations of text. In this approach, a user's entire tweet history (or a sample of their posts) is aggregated and

transformed into a high-dimensional vector embedding. Sentence-transformers or other language models (like BERT-based models[2]) can encode the semantic content of tweets into dense vectors. Each user can be represented by a vector that captures their overall opinion. Supervised models (such as neural networks or even simpler classifiers) can then be trained on these vectors to predict political leaning. Recent studies demonstrate the effectiveness of deep learning on this task: for example, researchers like Kim et al. (2025) have fine-tuned transformer models on political stance detection datasets. They obtained good results for inferring stance from text alone but incorporating external knowledge (e.g. prompting large language models like ChatGPT or domain-specific BERT variants) has further boosted performance on benchmark datasets. Overall, text-based methods can capture subtle indicators of ideology that network methods might miss. However, they require substantial training data with labels (users whose affiliation is known, perhaps via self-identification or following a known politician), and purely textual cues can sometimes be noisy or context-dependent.

The current state-of-the-art in political affiliation inference often combines network and content features, taking advantage of both what users say and with whom they interact. Research indicates that such fusion achieves the highest accuracy. For example, Aldayel and Magdy (2019) showed that a model combining users' tweet content with their network interactions outperformed content-only or network-only models. More recently, graph-based learning techniques have been applied: Peng et al. (2024) built a graph neural network over a bipartite graph[3] of users and tweets, and integrated BERT embeddings of tweet text as node features; it produced much better results than a text-only BERT model on the same data. Another notable approach is Retweet-BERT by Jiang et al. (2023), which exemplifies the fusion of diffusion patterns and language. Retweet-BERT uses features from the retweet network alongside the language in users' profiles, under the assumption that both network ties and self-described interests reflect ideology.

Underlying many of these methods is a concept that we explore deeply in this project: ideological contagion or diffusion. If one treats political orientation as an attribute that can spread or reinforce through social ties, algorithms like label propagation essentially simulate that spread until the network reaches a labeled equilibrium. This approach has been validated by the observation of strong clustering of political discourse on X (often described as echo chambers[4]). Some works, have also taken a longitudinal view –examining how these networks and inferred leanings change over multiple years or election cycles–. While users' core affiliations tend to be stable,

[2]BERT is a deep learning model based on transformers and it stands for Bidirectional Encoder Representations from Transformers Abas et al. (2022). A transformer is a particular neural network architecture that uses an "attention" mechanism, which is a way for the system to learn which parts of inputs are more relevant for which other parts of input, and correspondingly to which parts of output as well.

[3]A bipartite graph is a graph where the vertices can be divided into two disjoint sets such that every edge connects a vertex from one set to a vertex in the other set and no edges exist within the same set.

[4]In the media and social networks, an echo chamber is an environment in which people encounter beliefs that reinforce their preexisting thoughts by communication and repetition inside a closed system and away from counterarguments.

the network structure can change over time (for instance, new partisan influencers emerge). Multi-year analyses of X's political networks have documented increasing segregation of information sources and the persistence of homophily-driven clusters. This suggests that methodologies like those above remain robust over time, though researchers must be mindful of temporal dynamics (e.g. a model trained on 2016 election tweets may need adaptation for 2020 terms and expressions).

Overall, the state-of-the-art approaches to inferring political affiliation synthesize social graph signals, textual content, and advanced machine learning, achieving high accuracy in mapping online users onto the political spectrum. Ongoing research continues to refine these models, for example by applying community detection to find latent ideological groupings without prior labels. These efforts contribute to our understanding of online opinion clusters, information diffusion, polarization and the structure of political dialogue on social media with great detail and across extended time periods.

Chapter 4

Data Retrieval

Chapter Summary: This chapter details the acquisition of the datasets used in this research, covering both the original data from the 2016 and 2020 U.S. presidential elections and the newly collected data for the 2024 election. It explains how the initial datasets were obtained via Twitter's API and the subsequent pivot to a custom web scraping solution for 2024 due to recent platform policy changes. The chapter elaborates on the technical challenges encountered during the scraping process, such as handling dynamic content, platform restrictions, and the implementation of strategies to ensure robust and continuous data collection. Finally, it provides a comprehensive description of the filtered datasets for each election year.

4.1. Original Data: 2016 and 2020

The data used in this research was generously provided by Professor Rafael Caballero Roldán and his research team, who originally collected it as part of their own academic work. It was collected using Twitter's API restricting the search to tweets from the election weeks that mentioned the usernames of the presidential candidates.

This dataset consists of tweets published from one week before the election day to two days after, covering both the 2016 U.S. presidential elections (Donald Trump vs. Hillary Clinton) and the 2020 elections (Donald Trump vs. Joe Biden).

The data includes information from both individual tweets and the users that published them. For both 2016 and 2020, each tweet record primarily includes tweet ID and user ID, date of creation, whether it is a retweet, and in such case, the original user and the full text of the tweet. The volume of tweets is approximately 2.8 million tweets from 2016 and 1.6 million tweets from 2020. We also have information about the approximately 200,000 common users who were active in both electoral cycles, including the user ID, screen name, number of followers, number of tweets and retweets, location and other available metadata.

4.2. Scraping Overview

In order to extend the research to the upcoming 2024 U.S. presidential elections (Donald Trump vs. Kamala Harris), we aimed to replicate the same data collection strategy used for 2016 and 2020. The goal was to obtain tweets covering the week before and four days after election day (to leave a safe margin for errors), and to do so specifically from the users who had already posted in the previous elections. This approach would allow us to preserve the longitudinal nature of the dataset and analyze the evolution of user behavior over time.

However, due to the recent policy changes introduced by Twitter (now X), the use of its official API was no longer a viable option because of budget constraints. As a result, the data collection process was carried out through a custom web scraping system. The scraper extracts the user screen name, date of publication and textual content of the tweet, excluding images, but is capable of encoding emojis as part of the text.

The scraper was implemented using Jupyter Notebooks and developed in Python, making extensive use of Polars and Pandas for data handling, Selenium for browser automation, and ChromeDriver for navigating user profiles. In essence, the scraper iterates over the target users –which are the intersection of users that posted at least one tweet in the 2016 and in the 2020 elections–, visits their profile pages, and continuously scrolls down to load and extract their tweets within the desired date range. The speed and efficiency of this process largely depend on the frequency of publication of each user and the recency of their tweets: users who post infrequently and whose target tweets are recent can be processed significantly faster. To improve speed and performance, the code was optimized to run in headless mode.

In order for the process to work correctly, we needed to log into an account on the application. Without logging in, tweets would not appear in chronological order. These logins, along with cookie rejection and search initiation, were also handled automatically.

4.3. Technical Challenges

Throughout the scraping process, we encountered several technical challenges. One significant limitation imposed by the platform's design was that using a date range far in the past was highly inadvisable. Scrolling to older content could take an excessive amount of time, and at some point, X would stop displaying older tweets (even if they still existed). Interestingly, these tweets could sometimes be retrieved using the platform's date-filtered search feature, but even then, X only returned a sample, not the complete set of results. Consequently, the most reliable strategy to retrieve the 2024 data was to scrape tweets by scrolling down chronologically, in particular, the scraping started one week after the presidential election in order to improve performance. This would have been impossible in most cases if we had tried to retrieve tweets from 2016 or 2020 using this method, because we were scraping them years later and many of these users were highly active. Nevertheless, since we already had this data it was not a problem and if we had needed it we would have

implemented it using the advance search functionality mentioned before.

Another issue was that we had to carefully time the scroll-down action to ensure all tweets on the current page had been fully loaded and stored. To do this, we tracked the text of the last visible tweet during each scroll and waited until a new tweet with text appears. However, this approach presented challenges. For example, some users occasionally posted the same tweet multiple times, which made relying on tweet text unreliable. Moreover, tweet IDs were dynamic –they regenerated when scrolling– so we couldn't use them as stable references. Instead, we considered two tweets to be identical if they shared the same text and date. Complicating matters further, some users only posted videos or images. Since our scraper ignored these types of posts, the tracker for the last tweet read wouldn't update, potentially causing the scraper to misinterpret the page as unchanged. To address this, we implemented additional scrolls to account for such edge cases. Another issue was that scroll-down operations weren't disjoint: tweets would often be repeated across scrolls because the application retained previously seen content. To prevent duplicate processing, we kept a set of (text, date) pairs for the current user being processed. This structure allowed for efficient duplicate detection.

In addition, when the text of a tweet included different elements like hashtags, mentions, or emojis, the application split them into separate containers. We had to extract and concatenate all these parts to reconstruct the full tweet text accurately.

An important note is that some of the users we attempted to scrape had become private, had been banned by the platform, had deleted their accounts, or simply had no remaining posts because they had erased them. These cases had to be detected programmatically so the scraper could skip them and proceed to the next user.

The scraping process also required handling various platform-specific details. For instance, we had to automatically dismiss prompts encouraging users to subscribe to the paid version of X and accept warnings like those that informed of when a profile included potentially sensitive content.

However, one of the most significant challenge was that X frequently detected unusual activity and responded by displaying a "something went wrong" message, temporarily blocking the account being used to scrape. To overcome this obstacle, we redesigned the system completely to base it on multiple X accounts. We created approximately 30 accounts and the scraper would automatically cycle through these accounts, logging in and out as needed to avoid detection and bypass temporary blocks, while also inserting short random delays to make the behavior appear more human-like and unpredictable. Using multiple accounts also helped address an additional constraint that frequently appeared during the initial stages of data collection: it allowed us to overcome the daily limit on the number of tweets viewable from a single free account.

With these improvements, the program was able to operate continuously for days before encountering an unavoidable driver failure related to extended runtime of the navigation system. Upon such failure, the system would alert with an audible notification, automatically save all current progress, and enable seamless resumption of operations from the point of interruption by simply restarting the script.

The scraping process was executed over several weeks, using multiple devices simultaneously, all operating within the same collection window. In addition, the

system was designed to allow safe interruption and resumption at any point. This was achieved through the use of local copies of the data and the MongoDB Atlas server, which enabled secure concurrent work across different machines. The cloud database allowed us to run several instances of the program in parallel without interference, as the names of the already processed users were uploaded to the cloud and checked before processing a new user.

4.4. Data Description

After preprocessing the raw data –which included removing empty strings, blank tweets, tweets outside the selected date range, and filtering for posts written in English– the final dataset consisted of approximately 1.5 million tweets from 17,600 users. The reduction in the number of users retrieved compared to those initially searched is due to the fact that, unfortunately, many of the accounts were no longer active, had been banned or had become private during the past years.

To ensure consistency across all electoral cycles, we applied the same filtering criteria to the 2024 dataset as was originally applied to the 2016 and 2020 datasets provided to us. Specifically, in the original datasets for 2016 and 2020, tweets were pre-filtered to include only those mentioning the username of any of the candidates running for president. Following this approach, we retained for 2024 only those tweets that mentioned exactly one of the following usernames: `@realDonaldTrump`, `@KamalaHarris`, or `@JoeBiden`.

In addition, to accurately assign sentiment and political orientation within each tweet, we required that each tweet mention only a single political party (i.e., only one candidate username per tweet). We also applied further filters to ensure that all tweets fell within the same temporal window across the three elections, that tweets were written in English and that only users who appeared in all three electoral cycles were retained (i.e., the intersection of users from 2016, 2020, and 2024).

After applying all these filters, the resulting dataset sizes were as follows:

- 36,089 tweets for 2016
- 18,106 tweets for 2020
- 9,415 tweets for 2024

This filtering process naturally led to a substantial reduction in the total number of tweets, with the primary cause being the restriction that each tweet must mention only one candidate for coherence and simplification reasons explained in 5.1.

Out of curiosity and without relevance to the current project, in order to obtain a more realistic estimate of the total volume of political tweets collected in our 2024 dataset (which could be valuable for future research projects), we explored alternative methods. Initially, we experimented with several text classification models available on Hugging Face, but these models did not yield satisfactory results in distinguishing political content in a social networks context. As an alternative, we turned to Named Entity Recognition (NER) models, which allowed us to identify

and extract all named entities classified as persons within the tweets. From the NER results, we selected the 500 most frequently mentioned entities and manually filtered them to retain only political Figures. Using this list of political names, we retrieved all tweets containing at least one of these entities. This approach yielded a total of at least 640,000 political tweets from the 2024 election, showing that we had obtained a decent dataset for political analysis and research.

Chapter 5

From tweets to political opinions

Chapter Summary: This chapter is the main chapter of the thesis. It details the methodological framework for inferring political opinions from user data, starting with an acknowledgement of inherent platform biases and outlining key logical assumptions, such as treating candidates as binary opposites. It then thoroughly describes the chosen classification method, which involves representing users as vector embeddings of their concatenated tweets, establishing graph connections based on common tweets and cosine similarity, and subsequently propagating political labels from a small seed set of manually classified users. The chapter explains the three sequential contagion mechanisms created and presents the cumulative classification results for 2016, 2020, and 2024. Finally, it discusses preliminary attempts with other methods, analyzes the impact of heuristic thresholds and propagation order, and validates the chosen approach through performance analysis.

5.1. Framework assumptions

To begin with, although our work does not aim to perform any political or sociological analysis, it is important to note that the social network under study brings its own initial bias. Access to the platform is not uniform across the population, and it has inherent user profiles and recommendation algorithms (Glaser et al., 2024) as well as the possibility of manipulation through bots (Binns et al., 2023), among other factors.

Beyond this initial platform bias, we make several logical assumptions by restricting our analysis to two candidates, treating each one as the negation of the other, and applying the basic principles of binary logic. These assumptions are detailed below:

1. **Identity ($A = A$).** We assume that, given a user, a candidate, and an election year, that user holds exactly one opinion about that candidate in that election. In practice, opinions may evolve over the course of the campaign, but to simplify our study we ignore temporal variation, which would require a

chronological analysis left for future work. In other words, we do not take into account the exact timestamp of each tweet and instead associate each user with a set of tweets in the election under consideration.

2. **Principle of Non-Contradiction ($\neg(A \wedge \neg A)$).** No user supports both candidates simultaneously. While we initially considered allowing for neutral users –those who might express support for both or for neither– the proportion of such cases was under 1 %, so they were excluded from our analysis. Moreover, previous years' data confirmed that users rarely express contradictory support or opposition to both candidates.

3. **Principle of Excluded Middle ($A \vee \neg A$).** We assume that for any user in a given election, supporting one candidate implies not supporting the other. For example, in 2020 we treat Biden as the logical negation of Trump. Consequently, criticizing one candidate is taken as implicit support for the other. We acknowledge that this introduces bias: some users may dislike both candidates and third-party candidates do exist. Notably, in 2016 third-party candidates collectively received over 5 % of the total vote (Federal Election Commission, 2017).

On our side, we also introduce the following methodological biases:

- By selecting only those users who participated in the 2016, 2020, and 2024 elections, we inherently exclude younger voters who did not meet the age requirement in earlier years.
- In data collection, we consider only tweets that explicitly mention the official names of the candidates, omitting informal or colloquial references.

In addition, a further bias arises from the granularity chosen when determining a user's political opinion from their tweets:

1. **Tweet-level analysis.** Analyze each tweet separately (e.g., via sentiment analysis), then aggregate the per-tweet opinions into a single user-level opinion using a mathematical function (mean, mode, etc.).

2. **User-level aggregation.** Concatenate all the user's tweets into a single text and classify the combined text. This is the method we employ, having labeled concatenated tweet vectors.

In the tweet-level approach, it is crucial to clarify which candidate each tweet refers to. To this end, we impose an additional bias by selecting only those tweets that mention exactly one candidate by name, thereby avoiding tweets that could ambiguously refer to multiple targets.

Finally, another source of bias we introduce is related to the treatment of ambiguous messages. Despite our efforts to assign a clear political stance to each tweet, some messages –such as *"Raw footage of mailroom in post office here in Miami Dade. Source revealed 'mail-in ballots are within these piled up in bins on the floor. Mail*

has been sitting for over a week!"'– were associated with both candidates. Tweets like this one were shared by supporters of the Democrat and Republican parties in the 2024 election, each interpreting the situation as evidence that the opposing side was attempting to manipulate the electoral process. Such cases illustrate how the same content can have varying political interpretations depending on the user's alignment, making strict political opinion attribution a potential source of misclassification and bias. Fortunately, such messages represent rare exceptions within the overall dataset.

5.2. Classification Method: Vectors, Connections and Propagation

The objective of this stage is to classify users as either Republican or Democrat based on their tweets regarding both political parties. The final chosen methodology relies on encoding users as vectors that represent their opinions, connecting users with similar opinions, and then propagating labels from a small set of manually classified users across the network. For each election year, we construct a specific graph where nodes represent the users.

5.2.1. User Representation through Vectors

Each user is encoded as six separate entities, corresponding to their opinions on the Democratic and Republican candidates across the three electoral cycles:

- `XXX_16Dem` : opinion of user XXX on Democrats in 2016
- `XXX_16Rep` : opinion of user XXX on Republicans in 2016
- `XXX_20Dem`, `XXX_20Rep`, `XXX_24Dem`, `XXX_24Rep` : analogous for 2020 and 2024

For each of these entities, we concatenate all the tweets (from that year and referring exclusively to the corresponding candidate) into a single large text, separating individual tweets with a period.

We then use the `sentence-transformers` Python library to compute vector embeddings for each text using Sentence Transformer models Reimers and Aarsen (2025). This results in a numerical representation of each entity in a continuous vector space with more than 16,000 dimensions, each normalized between -1 and 1.

5.2.2. Graph Connections

For each year, we build a single graph where nodes correspond to users (`XXX_AA`), without the political extension of Democrat or Republican opinion. Thus, we have three graphs, one for each election considered.

Edges between nodes are established according to two criteria:

- **Common Tweets Connection**: We connect users who have shared at least one tweet in common, establishing a *Common* edge between them. This has proven to be very useful because of the great amount of retweets among users of the same party.

- **Similarity Connection**: We compute cosine similarity (1 - cosine distance) between vectors of the same type (Democrat-to-Democrat or Republican-to-Republican) within the same year. A *Similarity* edge is created between two users if their cosine similarity is equal to or greater than 0.85.

  Since each user has both a Democratic and Republican vector, it is possible that a pair of users (`XXX_AA` and `YYY_AA`) are connected twice: once through their Democratic similarity and once through their Republican similarity. In that case, we define the final similarity between them as the average of both values.

  This resulting similarity measure ranges from 0 to 1, is reflexive (maximum similarity of a user with themselves), and symmetric (the similarity from `XXX` to `YYY` is the same as from `YYY` to `XXX`) but it does not exhibit the triangle inequality property and therefore the cosine similarity is not a true distance metric.

The outcome of this process is a graph for each year where nodes represent users and edges represent a meaningful relationship, either through content interaction (shared tweets) or through opinion similarity (taking into account both partisan perspectives). This graph serves as the foundational structure for the final classification process based on label propagation.

5.2.3. Label Propagation

Once the graphs were built, the next step was to classify users as either Democrats or Republicans. This process began with the manual classification of a set of users, later extended through label propagation across the graph. These graphs and processes were implemented using `neo4j` and executed through `Cypher` queries.

To facilitate the initial labeling, we employed the natural language processing Python library `flair`. Although `flair` generally performs poorly when classifying individual tweets as either Democratic or Republican, if we instead evaluate all aggregated tweets from a user and under certain specific conditions its performance becomes highly reliable. In particular, when a user has published at least 8 tweets, and `flair` provides a prediction with a confidence higher than 85 %, the accuracy in classifying users exceeds 97 %, as verified through manual inspection. Therefore, under these circumstances, `flair` was used to assist in creating a larger and trustworthy seed set of classified users.

Combining manual labeling and this assisted classification method, we obtained an initial seed of 900 users labeled as either Democrats or Republicans.

These initial labels were then propagated throughout the graph, allowing unlabeled users to inherit the labels of their neighbors in successive stages, applying three contagion mechanisms in sequence:

- **Common tweets contagion**: An unlabeled user adopts a label if a sufficiently large proportion of their neighbors with whom they share tweets belong to the same party. In particular, the user adopts a label if more than 15 % of their common tweets neighbors belong to that party, provided that less than 10 % belong to the opposing party.
- **Similarity-based contagion**: An unlabeled user adopts a label based on the average similarity with their already labeled neighbors. Specifically, a user is classified as Republican if their average similarity with Republicans exceeds 0.2, while their average similarity with Democrats is below 0.1 or nonexistent. Conversely, a user is classified as Democrat if their average similarity with Democrats exceeds 0.2, while their average similarity with Republicans is below 0.1 or nonexistent.
- **Majority-based contagion**: An unlabeled user adopts the majority label among all their direct neighbors, regardless of the type of connection (similarity or common tweets). If more than 90 % of a user's neighbors belong to a single party, the user adopts that majority label, idea supported by Auletta et al. (2018).

These contagion mechanisms were applied iteratively, progressively coloring the graph until reaching a fixed point where no further users could be labeled.

It is important to note that, even after the propagation process was completed, some nodes remained unclassified. For example, the isolated nodes without similarity or common tweet edges to any other user.

The propagation process was applied sequentially in three distinct stages. First, all rounds of propagation based on *common tweets* connections were executed until no new users could be labeled. Once this stage reached a stable point, the same iterative procedure was applied using *similarity-based* propagation. Finally, a last propagation stage was performed based on the *majority* rule, again repeated until the graph stabilized and no additional labels could be assigned.

Although, in theory, repeating the entire propagation sequence could lead to new classifications, in practice, this was not the case in our data. After completing the three stages of propagation in the described order, the graphs reached a fixed point where no further labels could be propagated.

Figure 5.1 shows the progression of the graph of the year 2024 with only the manually labeled nodes colored at first and the final picture after all the propagation processes are applied. The figure shows that most of the unclassified nodes are isolated nodes.

The tables 5.1, 5.2 and 5.3 summarize the final classification cumulative numbers resulting from this methodology for each year. In general, the first contagion method to be applied, in this case the common tweets contagion, is the one that usually makes the most extensive contagion.

The Figure 5.2 provides a visual example of the label propagation process, illustrating the case of user `@lulubellehen` of 2016, positioned at the center of the graph and connected to her 26 neighboring users. The sequence of panels from Figure 1 to Figure 4 shows the progression of the graph coloring as label propagation unfolds.

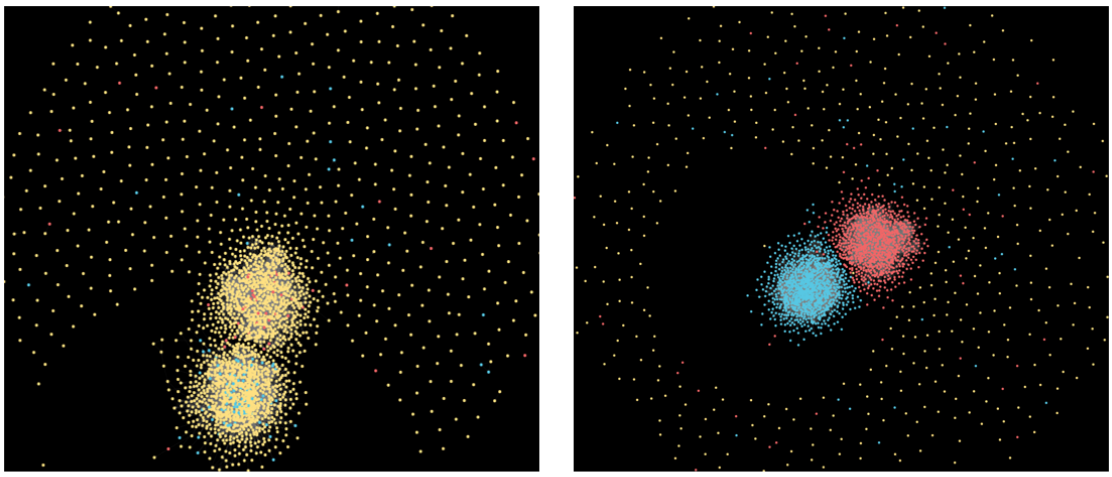

Figure 5.1: Graph of 2024 in the initial stage (pre-contagion) in Figure 1 and in the final stage (after applying the three contagion methods) in Figure 2. The background is colored in black to highlight isolated nodes. Democrat nodes are painted in blue, Republican nodes in red and unclassified nodes in yellow. Source: Own elaboration.

Table 5.1: Contagion Process Cumulative Results - Year 2016

Process Stage	**DEM Nodes**	**REP Nodes**	**Unclassified Nodes**
Initial (Pre-Contagion)	201	252	2196
Post-Common Contagion	959	899	791
Post-Similar Contagion	973	914	762
Final (Post-Majority)	992	923	734

Table 5.2: Contagion Process Cumulative Results - Year 2020

Process Stage	**DEM Nodes**	**REP Nodes**	**Unclassified Nodes**
Initial (Pre-Contagion)	142	119	2396
Post-Common Contagion	909	925	823
Post-Similar Contagion	910	925	822
Final (Post-Majority)	919	931	807

Table 5.3: Contagion Process Cumulative Results - Year 2024

Process Stage	**DEM Nodes**	**REP Nodes**	**Unclassified Nodes**
Initial (Pre-Contagion)	125	49	2483
Post-Common Contagion	1198	955	504
Post-Similar Contagion	1204	962	491
Final (Post-Majority)	1211	964	482

In Figure 1, the original state of the graph is shown, with only manually labeled nodes: red indicates Republican users, and blue indicates Democrat users. Figure 2 shows the result after applying the *common tweets* propagation step. At this point,

all of `@lulubellehen`'s neighbors have been labeled through propagation from other nodes. In Figure 3, the graph is updated after the *similarity-based* propagation step. In this case, `@lulubellehen` remains unlabeled. Her average similarity to Republican neighbors is 0.05496, and to her sole Democrat neighbor, 0.104. Since neither of these values meets the thresholds required for propagation, her node is not classified during this stage. Finally, Figure 4 displays the outcome after the *majority-based* propagation step. At this stage, `@lulubellehen` is successfully labeled as Republican, due to the overwhelming number of Republican neighbors in her immediate network. This result confirms the reliability of the method: upon inspection of her actual account, `@lulubellehen` self-identifies as a proud "MAGA Mother" and refers to Donald Trump as her "Warrior President".

Figure 5.2: Set of Figures illustrating the label propagation process of user `@lulubellehen` of 2016. For the sake of clarity, only the edges that connect directly to this central user are displayed; the edges between other users are omitted. Each node is labeled with the corresponding username, and each edge is annotated with the type of connection it represents (*SIMILAR* or *COMMON*). Republican users are colored in red, Democrats in blue and unlabelled nodes in light brown. Source: Own elaboration.

5.3. Methodological Overview and Validation

5.3.1. Preliminary Attempts and Design Choices of the Classification Method

Before arriving to this method we initially explored other classification approaches based on sentiment analysis and NLP. The main alternative we considered used the `flair` and `textblob` Python libraries. We manually labeled a sample of 300 tweets, independently evaluated by three human annotators. The agreement between annotators was measured using Cohen's Kappa coefficient. Each tweet was classified with a score of -1, 0, or 1, corresponding respectively to a negative, neutral, or positive sentiment towards the political party mentioned in the tweet. However, the results obtained were unsatisfactory, with Cohen's Kappa values ranging between 0.2 and 0.5. These poor results are likely explained by the particular characteristics of the language used in the dataset. The political discourse on social media

presents multiple challenges for automatic classifiers that have not been extensively fine-tuned –a process that would require a far larger set of manually labeled samples–. Particularly, the tweets analyzed are characterized by informal social media language, full of recent and unusual expressions, misspellings, hashtags, emojis, and abbreviations. In addition, they are full of political references[1] that are often counterintuitive or ironic, making them difficult for sentiment models to interpret correctly.

After exploring those other options, we finally obtained the best performance results with the main method presented in the previous Section 5.2. However, some considerations must be made regarding the method described above. In particular, this section discusses the choices made regarding thresholds and the order of application for the three label propagation methods, as well as method variants that show better performance for this specific problem. These selected variant will be applied to the data, with results analyzed in Chapter 6.

The numerical thresholds presented in the method description –such as the lower bound of 0.85 for creating a *similarity* edge, or the percentage thresholds used in the three propagation mechanisms– were selected heuristically. These thresholds act as hyperparameters of the method and the stricter the threshold, the higher the expected classification accuracy, but the lower the number of nodes that will meet the condition and be labeled. Conversely, relaxing the thresholds allows for broader coverage at the cost of reduced reliability. To better understand this trade-off, one could construct a plot showing, at each threshold level, the relationship between how many users are labeled and what proportion of them are correctly classified. This analysis would allow tuning the method based on the specific needs of the problem: whether one prioritizes data quantity or label accuracy.

Regarding the order in which the propagation mechanisms are applied, while the *majority-based* propagation must necessarily be positioned last –as it relies entirely on the previous labels assigned within the graph– the order between the *common tweets* and *similarity-based* propagation methods was not predetermined. Empirical analysis showed that the execution order of these two steps had a very limited impact on the final classification of users. Specifically, in 2016, only 53 users received different final labels when comparing the two sequences: (*common tweets* $\rightarrow$ *similarity* $\rightarrow$ *majority*) versus (*similarity* $\rightarrow$ *common tweets* $\rightarrow$ *majority*). In 2020, the difference was reduced to a single user, while in 2024 the discrepancy involved only 21 users.

[1] Some illustrative examples include the case where Hillary Clinton referred to Trump supporters as "deplorable" for his racist and sexist behavior. In response, many of Trump's supporters began proudly adopting the label "Deplorable for Trump" and "Proud Deplorable", creating buttons and merchandise with the term. Although the word "deplorable" is negative, its appearance in these tweets actually indicates a positive attitude towards Trump. A similar event occurred when Clinton was speaking about raising taxes on the wealthy, Trump interrupted her, saying, "Such a nasty woman." The remark was widely criticized and quickly became a feminist slogan. Many women embraced the phrase "Nasty Woman" as a symbol of empowerment. Another example of the 2024 election involves a squirrel named Peanut, which became a political symbol after Republicans claimed that Democrats euthanized it. Consequently, every appearance of the word "Peanut" in tweets tended to convey a positive meaning for Republicans. These cases highlight the limitations of using generic sentiment classifiers in this context.

5.3.2. Performance Analysis and Comparison

We also tested some variations of our method and evaluated their performance. To do so, we manually classified 150 users based on their tweets from a specific election year. These users were selected to include 50 from each of the three propagation types: *common*, *similarity*, and *majority*. Their manual labels (Republican or Democrat) were then compared with the labels assigned by the full classification method described earlier.

The results were as follows:

Accuracy = 0.8933, **Recall** = 0.8765, **Cohen's Kappa** = 0.7862.

The confusion matrix is shown in Figure 5.3.

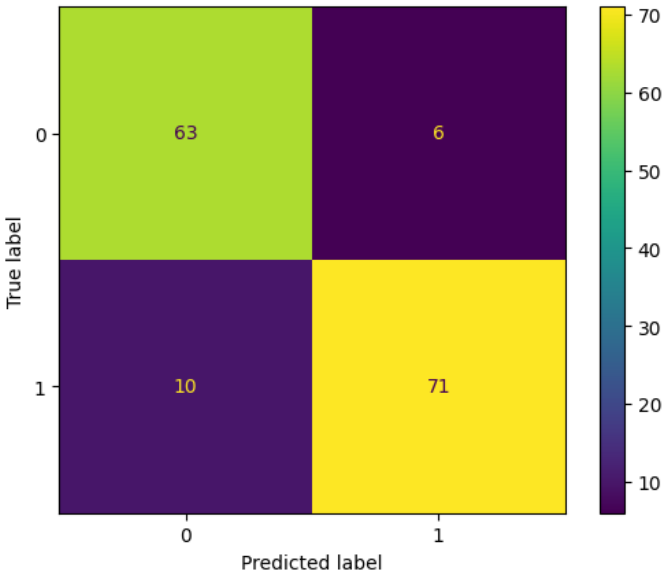

Figure 5.3: Confusion matrix obtained from a sample evaluation of the classification method that uses the three propagation types. The Republican label is encoded as 0 and the Democrat label as 1. Source: Own elaboration.

Upon analyzing the discrepancies, we found that all of them came from the *similarity* and *majority* propagation steps. In particular, 75 % of the errors came from nodes labeled through similarity propagation and 25 % through majority. Therefore, the *common* label propagation achieved perfect performance on its subset, correctly labeling all 28 Republicans and 22 Democrats it classified.

Summarizing, when the propagation types are evaluated separately:

- **Similarity propagation**: Accuracy = 0.76, Recall = 0.6897, Kappa = 0.5261
- **Common propagation**: Accuracy = 1.0, Recall = 1.0, Kappa = 1.0
- **Majority propagation**: Accuracy = 0.92, Recall = 0.9667, Kappa = 0.8305

Based on these results, for the purposes of this study, where we aim for statistically significant outcomes, we chose to use only the *common tweets* propagation

method, as it consistently outperformed the others in terms of precision. However, we acknowledge that the other propagation strategies still contribute additional node coverage and show reasonably good performance. In other contexts, the full method or its variants could prove more useful depending on the specific goals and trade-offs of each case.

Chapter 6

Results

Chapter Summary: This chapter presents the key findings derived from the analysis of political opinions. The first two sections focus on the subset of users classified through the highly accurate common tweets label propagation method and the third section does a global analysis on the filtered datasets obtained from the scraping process. The first section focuses on *who* are the classified users, detailing the distribution of classified Democrat and Republican users across the 2016, 2020, and 2024 election years. The second section covers *what* political content is being discussed, it analyzes content production, examining how each party's users distribute their attention between their own candidate and the opposition, revealing distinct messaging strategies across election cycles. Finally, the last section answers *how* tweet are composed over time by providing descriptive statistics on tweet features, illustrating changes in the usage of emojis, hashtags, mentions, and URLs across the three elections, reflecting broader trends in social media communication.

For all results reported as statistically significant, we include a Summary Table 6.2 at the end of the section showing the corresponding p-values and confidence intervals for each case.

6.1. User Demographics

The results presented in this section and in the following are based on a dataset of 5,837 user-year pairs. Each pair represents a unique user in a specific election year and corresponds to a node in the final classification graphs. While it is possible that the same individual appears in multiple years, we refer to each user-year pair simply as a "user" for clarity.

All users included in this dataset were labeled either manually or through the *common tweets* label propagation method which classifies a user if a sufficiently large proportion of their neighbors with whom they share tweets belong to the same political party. This propagation rule has demonstrated the highest classification accuracy and is therefore used as the foundation for our analyses.

We restrict our results to this subset because it represents a highly reliable sam-

ple, allowing for meaningful and robust statistical analysis. On the downside, not all users fulfilled the criteria to be assigned a label using this method, which means that the total number of users differs across election years. As a result, one cannot directly infer voter switching from differences in group sizes across years[1]. Instead, we focus on analyzing the aggregate behavior of the classified users for each year. This is the trade-off we accept in exchange for a more accurate classification approach.

The distribution of users by year and party affiliation is shown in the following table:

Year	Democrats	Republicans
2016	955	899
2020	909	923
2024	1198	953

An analysis of the proportion of Democrat and Republican users over time reveals a significant increase in Democrat representation in 2024 compared to 2016 and 2020. This change in party distribution is statistically significant, as confirmed by both the chi-squared test and two-proportion Z-tests conducted between years[2]. While these results do not reflect the actual election outcomes, this is not unexpected given that our dataset is relatively small and subject to several biases, such as those discussed in Section 5.1.

Additional significant statistics of users for specific years include the following: in 2016, Republican users had fewer followers on average than their Democrat counterparts, and the number of verified Republican accounts was also lower compared to Democrats.

6.2. Content Focus

We now turn to analyzing the content produced by users, in order to identify the typical profile associated with each party in each election year. In particular, we examine how users distribute their attention between the Democratic and Republican candidates. This can show possible changes in communication strategies across election cycles.

For each user, we compute the proportion of their tweets that mention Democratic or Republican candidates. Aggregating these proportions by party affiliation and

[1]Nevertheless, as a side note, we can report that using the original combined method with common tweets, similarity, and majority propagation strategies –which offers lower accuracy– we found that party switching was marginal. Only approximately 30 users appeared to switch affiliations, a number that would likely be even smaller when accounting for the potential error margin of the method.

[2]The chi-squared test assesses whether there is a significant association between two categorical variables (in this case, year and political alignment), while the two-proportion Z-test evaluates whether the difference in proportions of users from a specific party between two years is statistically significant (done for every combination of two years). Although the Z-test is a parametric test that assumes normality, it can still be applied to large samples due to the Central Limit Theorem, which ensures that the sampling distribution of the mean approximates normality regardless of the underlying distribution.

election year allows us to visualize the average focus of Democrat and Republican voters on each party's candidate. Figure 6.1 presents the results of this analysis at the user level.

We repeated the same analysis at the tweet level, aggregating all tweets from all users per group, rather than averaging individual user proportions. In theory, these two approaches do not necessarily yield the same result due to the non-linearity of averages –one is based on the average of user-level proportions, while the other computes the overall mean across all tweets–. However, in practice, the resulting graphs are very similar, as shown in Figure 6.2. The differences are minimal and do not alter the main conclusions.

One minor difference can be observed in the case of Democrat voters in 2020. When analyzing the tweet-level proportions, the percentage of tweets about Trump increases significantly compared to the user-level analysis. This is due to a few outlier users who posted a large volume of tweets about Trump. Their influence is prominent in the tweet-level analysis but is diluted when averaging at the user level.

By examining the user-level graphs, we can derive meaningful insights into the messaging strategies associated with each party. For instance, among Democrat voters, we observe that in both 2016 (Clinton vs. Trump) and 2020 (Biden vs. Trump), users split their attention relatively equally between their own candidate and the opposition. However, in 2024 (Harris vs. Trump), the communication strategy appears to shift. The emphasis is primarily on promoting Kamala Harris and highlighting the potential improvements her presidency could bring, while largely avoiding direct engagement with Donald Trump.

This contrasts with the strategy observed among Republican voters. Their communication consistently focuses on praising their own candidate. In 2016, the number of tweets about Trump was relatively low, likely due to his lack of prior political experience, and thus criticism was more strongly directed toward Hillary Clinton. However, starting in 2020, after Trump completed his first presidential term, Republican users increasingly centered their messaging around celebrating Trump, while minimizing mentions of Democratic candidates.

Another key result is that, on average, Republican users tweet more frequently than Democrat users –a difference confirmed by the Mann–Whitney U test[3]–. This trend is particularly pronounced in the years 2016 and 2020. However, in 2024, the pattern slightly reverses: Democrats tweet more on average than Republicans. This shift could be attributed to several factors, such as the dual effort to defend both Biden's re-election campaign and promote Kamala Harris as the new leading candidate. Another possible explanation is a reduced Republican presence on X, as some influential figures migrated to alternative platforms such as Truth Social.

[3]The Mann–Whitney U test is the non-parametric statistical test equivalent of the Student's t-test used to compare two independent samples. This test is appropriate because the dependent variable is measured on a continuous scale, the independent variable consists of two independent groups, observations within each group are independent, and the distributions of the variables deviate from normality.

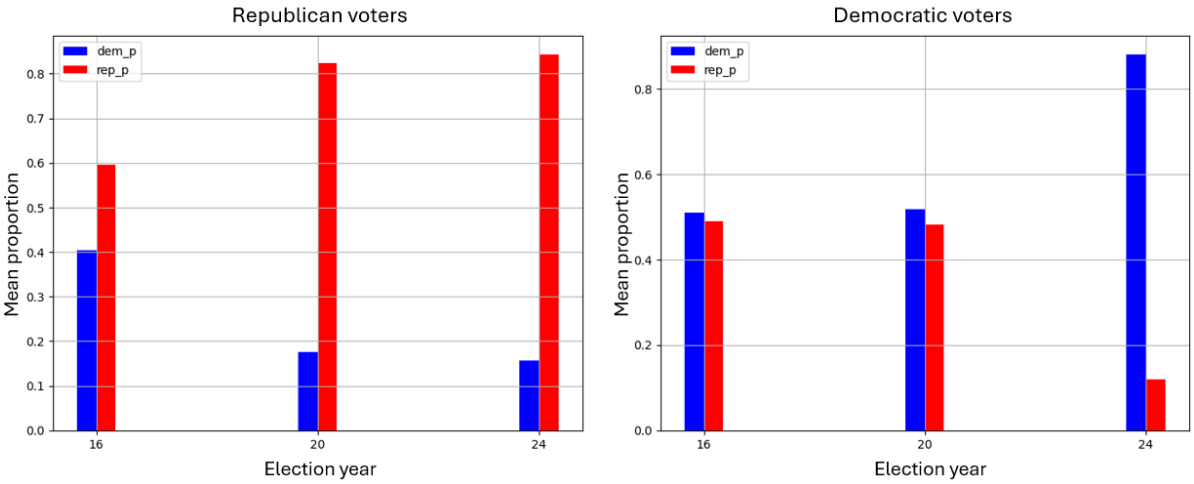

Figure 6.1: Distribution of attention to Democratic and Republican candidates by users, grouped by vote and year. Source: Own elaboration.

6.3. Tweet Composition

Finally, we present a set of descriptive statistics computed over the entire dataset of users obtained through web scraping, as introduced in Chapter 4. These statistics are not restricted to users labeled through common-based propagation, as they are not dependent on party affiliation.

These statistics reveal how the format of tweets have evolved across the three election cycles:

1. **Increase in emoji usage**: Compared to 2016, tweets in both 2020 and 2024 contain significantly more emojis per tweet. Additionally, 2020 shows a significantly higher average emoji usage than 2024.

2. **Decrease in hashtag usage**: There is a consistent decline in the average number of hashtags per tweet across the three elections.

3. **Stable user mention patterns**: The average number of user mentions per tweet increased significantly from 2016 to 2020, and then slightly decreased in 2024. Although the differences between 2020 and 2024 are small, they remain statistically significant.

4. **Sharp decline in URL usage**: The average number of URLs per tweet was significantly higher in 2016 and 2020 compared to 2024. The drastic drop in 2024 reflects changes in platform usage because it is no longer necessary to include the URL of a media file to be able to share it in a tweet, you can simply add the picture, video or audio directly.

To assess whether observed differences between the means computed for each year shown in Table 6.1 are statistically significant, we computed 95 % confidence intervals for the difference in means using a bootstrap resampling approach. All

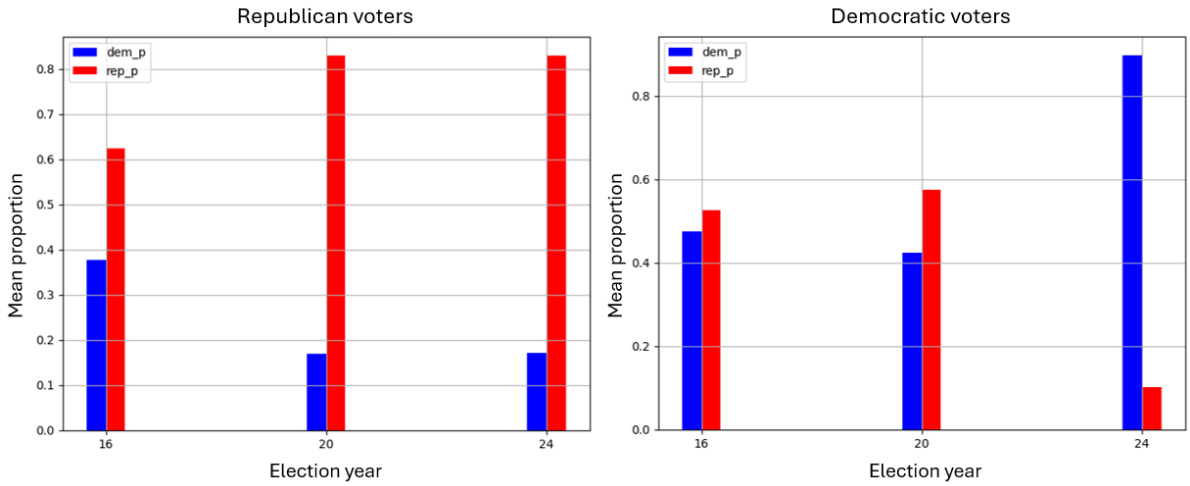

Figure 6.2: Distribution of attention to Democratic and Republican candidates by tweets, grouped by vote and year. Source: Own elaboration.

shown results are considered statistically significant since the corresponding interval of the differences does not include zero. In this case, we used confidence intervals instead of the Mann–Whitney U test because the medians were consistently zero and there were many tied values. Since the test relies on ranking differences between groups, the lack of variability and the high number of ties made it ineffective for detecting any meaningful difference.

Table 6.1: Average Tweet Features per Election Year

Feature	**2016**	**2020**	**2024**
Emojis per tweet	0.136	0.356	0.283
Hashtags per tweet	0.557	0.509	0.301
Mentions per tweet	1.831	1.889	1.650
URLs per tweet	0.442	0.495	0.023

These results highlight notable shifts in how political content is shared and formatted over time, reflecting evolving trends in social media and political communication.

Table 6.2: Summary of Statistical Findings on Political Discourse and User Characteristics

Finding	Statistical Test	Result (p-value / CI)
The proportion of Republicans/Democrats per year varies significantly	Chi-squared contingency test	$p = 0.0004$
Increase in Democrat representation from 2016 to 2024	Proportions Z-test	$p = 0.0081$
Increase in Democrat representation from 2020 to 2024	Proportions Z-test	$p = 0.0001$
In 2016, Republican users had fewer followers on average than their Democrat counterparts	Mann-Whitney U test	$p = 0.0003$
In 2016, the number of verified Republican accounts was also lower compared to Democrats	Mann-Whitney U test	$p = 0.0013$
Republicans tweet more than Democrats in general	Mann-Whitney U test	$p = 0.0003$
Republicans tweet more than Democrats in 2016	Mann-Whitney U test	$p < 0.0001$
Republicans tweet more than Democrats in 2020	Mann-Whitney U test	$p < 0.0001$
Republicans tweet less than Democrats in 2024	Mann-Whitney U test	$p < 0.0001$
Tweets in 2016 contained fewer emojis than in 2020	Confidence Interval	[-0.240, -0.200]
Tweets in 2016 contained fewer emojis than in 2024	Confidence Interval	[-0.162, -0.131]
Tweets in 2020 contained more emojis than in 2024	Confidence Interval	[0.052, 0.095]
More hashtags in 2016 than in 2020	Confidence Interval	[0.028, 0.068]
More hashtags in 2020 than in 2024	Confidence Interval	[0.184, 0.232]
More hashtags in 2016 than in 2024	Confidence Interval	[0.235, 0.275]
Fewer user mentions in 2016 than in 2020	Confidence Interval	[-0.084, -0.031]
More user mentions in 2016 than in 2024	Confidence Interval	[0.151, 0.210]
More user mentions in 2020 than in 2024	Confidence Interval	[0.203, 0.274]
Fewer URLs in 2016 than in 2020	Confidence Interval	[-0.064, -0.043]
More URLs in 2016 than in 2024	Confidence Interval	[0.413, 0.426]
More URLs in 2020 than in 2024	Confidence Interval	[0.463, 0.481]

Chapter 7

Conclusions and Future Work

Chapter Summary: This chapter summarizes the key achievements and insights derived from the project, beginning with the successful acquisition of a robust dataset of political tweets from X, despite significant web scraping challenges. It then highlights the effectiveness of the proposed classification method –particularly the common tweets label propagation variant– in accurately inferring user political affiliations. The chapter discusses the project's main findings from a political communication perspective. Finally, it outlines few ideas for future research.

7.1. Conclusions

A substantial portion of the project was dedicated to data acquisition, particularly due to the challenges posed by X during the web scraping process. Despite these obstacles, we successfully compiled a powerful dataset of at least 640,000 political tweets from the 2024 U.S. election. Although the dataset was later significantly filtered to ensure compatibility with the 2016 and 2020 datasets which only included tweets explicitly mentioning political candidates' usernames, enabling rigorous comparative analysis.

The proposed classification method was based on vector embeddings as nodes, edges created from cosine similarity and shared tweets, combined with three label propagation mechanisms:

- **Similarity-based contagion**: An unlabeled user adopts a label based on the average cosine similarity with their already labeled neighbors.
- **Common tweets contagion**: An unlabeled user adopts a label if a sufficiently large proportion of their neighbors with whom they share tweets belong to the same party.
- **Majority-based contagion**: An unlabeled user adopts the majority label among all their direct neighbors.

The method performed robustly, achieving an accuracy of 89.33 % and successfully classifying the majority of user nodes across all three years. Interestingly, a simplified variant that employed only the *common tweets* propagation achieved perfect accuracy (100 %) in inferring users' political affiliation and, despite producing less coverage, it was selected for the final analyses.

From a political communication perspective, our results reveal a clear evolution in voter behavior and campaign strategy. Over time, users increasingly focus on promoting their preferred candidate rather than criticizing the opponent. This trend is particularly pronounced in the 2024 election, where users of both parties primarily praise their own candidate. On the Republican side, the continuous presence of Donald Trump as a candidate has led to a noticeable increase in attention devoted to him across the three cycles, likely due to his rising political fame. We also observed temporal shifts in the number of tweets posted by each user. In 2016 and 2020, Republican users were generally more active than Democrats, while in 2024 this trend reversed.

Descriptive tweet statistics show a steady decline in the use of hashtags and a dramatic drop in the use of URLs, pointing to a transformation in the format and function of political discourse on the platform.

7.2. Future Work

There are multiple promising directions for extending this research:

- **Multimodal analysis**: Incorporating visual content –such as images, videos, and memes– could greatly enhance the representativeness of the dataset, as many tweets with meaningful political content were excluded due to their non-textual nature. Advances in image recognition and multimodal models make this a feasible and timely extension.

- **Graph-based enhancements**: The current method could be enriched through more advanced graph analysis techniques such as community detection, centrality-based influence modeling, or dynamic graph evolution over time.

- **Model fine-tuning**: Improving sentiment analysis performance through fine-tuned models trained specifically on political social media discourse would help address the limitations encountered with general-purpose classifiers.

- **Cross-platform generalization**: Applying the methodology to other social media platforms (e.g., Reddit, Facebook, Truth Social) could provide broader insights into the digital public sphere and capture politically active populations that have migrated away from X.

- **Cross-national applications**: Extending the model to study bipartisanship or multiparty dynamics in other countries would allow comparative political analysis across different electoral and media systems.

- **User profile analysis**: A deeper study of user profiles (e.g., age, location when available and biography text) could help better understand the characteristics of each political group.
- **Time series analysis**: Tracking how user opinions and network structures change day by day around key political events (e.g., debates, scandals, announcements) could give more detailed insights into the dynamics of polarization.
- **Interactive visualizations**: Developing interactive tools or dashboards for exploring the graphs and political alignments would make the findings more accessible and engaging for non-experts.

In summary, this project lays a strong methodology for the computational analysis of political discourse and voter alignment using social media data, while acknowledging future improvements and methodological extensions.

Chapter 8

Conclusiones y Trabajo Futuro

Resumen: Este capítulo resume los logros clave y las ideas derivadas del proyecto, comenzando con la exitosa adquisición de un robusto conjunto de datos de tuits políticos de X, a pesar de los importantes desafíos del *web scraping*. Luego, destaca la efectividad del método de clasificación propuesto –particularmente la variante de propagación de etiquetas de tuits comunes– para inferir con precisión las afiliaciones políticas de los usuarios. El capítulo discute los principales hallazgos del proyecto desde una perspectiva de comunicación política. Finalmente, se exponen algunas ideas para futuras investigaciones.

8.1. Conclusiones

Una parte sustancial del proyecto se dedicó a la adquisición de datos, particularmente debido a los desafíos planteados por X durante el proceso de *web scraping*. A pesar de estos obstáculos, recopilamos con éxito una potente base de datos de al menos 640.000 tuits políticos de las elecciones de EE. UU. de 2024. Aunque el conjunto de datos se filtró posteriormente de manera significativa para garantizar la compatibilidad con los conjuntos de datos de 2016 y 2020, que solo incluían tuits que mencionaban explícitamente los nombres de usuario de los candidatos políticos, lo que permitió un análisis comparativo riguroso.

El método de clasificación propuesto se basó en transformaciones vectoriales como nodos, aristas creadas a partir de la similitud del coseno y tuits compartidos, combinados con tres mecanismos de propagación de etiquetas:

- **Contagio basado en la similitud**: Un usuario no etiquetado adopta una etiqueta basándose en la similitud del coseno promedio con sus vecinos ya etiquetados.
- **Contagio de tuits comunes**: Un usuario no etiquetado adopta una etiqueta si una proporción suficientemente grande de sus vecinos con los que comparte tuits pertenece al mismo partido.
- **Contagio basado en la mayoría**: Un usuario no etiquetado adopta la eti-

queta mayoritaria entre todos sus vecinos directos.

El método funcionó de forma robusta, logrando una precisión del 89,33 % y clasificando con éxito a la mayoría de usuarios en los tres años. Curiosamente, una variante simplificada que empleó solo la propagación de tuits comunes logró una precisión perfecta (100 %) al inferir la afiliación política de los usuarios y, a pesar de ser capaz de etiquetar menos nodos, fue seleccionada para los análisis finales por su exactitud.

Desde una perspectiva de comunicación política, nuestros resultados revelan una clara evolución en el comportamiento de los votantes y en las estrategias de campaña. Con el tiempo, los usuarios se centran cada vez más en promocionar a su candidato preferido en lugar de criticar al oponente. Esta tendencia es particularmente pronunciada en las elecciones de 2024, donde los usuarios de ambos partidos principalmente elogian a su propio candidato. En el lado republicano, la presencia continua de Donald Trump como candidato ha llevado a un aumento notable en la atención que se le dedica a lo largo de los tres ciclos, probablemente debido a su creciente fama política. También observamos cambios temporales en el número de tuits publicados por cada usuario. En 2016 y 2020, los usuarios republicanos fueron generalmente más activos que los demócratas, mientras que en 2024 esta tendencia se invirtió.

Las estadísticas descriptivas de los tuits muestran un descenso constante en el uso de hashtags y una caída dramática en el uso de URLs, lo que apunta a una transformación en el formato del discurso político en la plataforma.

8.2. Trabajo Futuro

Existen múltiples direcciones prometedoras para extender esta investigación:

- **Análisis multimodal**: La incorporación de contenido visual –como imágenes, vídeos y memes– podría mejorar enormemente la representatividad del conjunto de datos, ya que muchos tuits con contenido político significativo fueron excluidos debido a su naturaleza no textual. Los avances en el reconocimiento de imágenes y los modelos multimodales hacen de esta una extensión factible y oportuna.
- **Mejoras basadas en grafos**: El método actual podría enriquecerse mediante técnicas de análisis de grafos más avanzadas, como la detección de comunidades, el modelado de influencia basado en la centralidad o la evolución dinámica de los grafos a lo largo del tiempo.
- **Ajuste del modelo**: Mejorar el rendimiento del análisis de sentimientos mediante modelos entrenados específicamente para el discurso político en redes sociales ayudaría a abordar las limitaciones encontradas con los clasificadores de propósito general.
- **Generalización multiplataforma**: La aplicación de la metodología a otras plataformas de redes sociales (por ejemplo, Reddit, Facebook, Truth Social) podría proporcionar información más amplia sobre la esfera pública digital y capturar poblaciones políticamente activas que han emigrado de X.

- **Aplicaciones transnacionales**: La extensión del modelo para estudiar el bipartidismo o el multipartidismo en otros países permitiría un análisis político comparativo entre diferentes sistemas electorales y mediáticos.
- **Análisis de perfiles de usuario**: Un estudio más profundo de los perfiles de usuario (por ejemplo, edad, ubicación y texto de la biografía) podría ayudar a comprender mejor las características de cada grupo político.
- **Análisis de series temporales**: El seguimiento de cómo las opiniones de los usuarios y las estructuras del grafo cambian día a día en torno a eventos políticos clave (por ejemplo, debates, escándalos, anuncios) podría proporcionar información más detallada sobre la dinámica de la polarización.
- **Visualizaciones interactivas**: El desarrollo de herramientas o paneles interactivos para explorar los grafos y las alineaciones políticas haría que los hallazgos fueran más accesibles y atractivos para los no expertos.

En resumen, este proyecto sienta una metodología sólida para el análisis computacional del discurso político y la alineación de votantes utilizando datos de redes sociales, al tiempo que reconoce futuras mejoras y extensiones metodológicas.

Bibliografía

Abas, A. R., Elhenawy, I., Zidan, M. and Othman, M. Bert-cnn: A deep learning model for detecting emotions from text. *Computers, Materials & Continua*, Vol. 71(2), 2022.

Akbik, A., Bergmann, T., Blythe, D., Fricke, K., Gossler, K., Schweter, S. and Vollgraf, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Association for Computational Linguistics, 2019.

Aldayel, A. and Magdy, W. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on human-computer interaction*, Vol. 3(CSCW), 1–20, 2019.

An, J., Cha, M., Gummadi, K., Crowcroft, J. and Quercia, D. Visualizing media bias through twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6, 2–5. 2012.

Auletta, V., Ferraioli, D., Greco, G. et al. Reasoning about consensus when opinions diffuse through majority dynamics. In *IJCAI*, 49–55. 2018.

Barberá, P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, Vol. 23(1), 76–91, 2015.

Binns, A., Bright, J. and Dubois, E. Not our kind of crowd! how partisan bias distorts perceptions of political bots on twitter (now x). *British Journal of Social Psychology*, Vol. 62(4), 1277–1297, 2023.

Bisgin, H., Agarwal, N. and Xu, X. A study of homophily on social media. *World Wide Web*, Vol. 15(2), 213–232, 2012.

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z. and Zimmer, M. Web scraping for research: Legal, ethical, institutional, and scientific considerations. *arXiv preprint arXiv:2410.23432*, 2024.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F. Predicting the political alignment of twitter users. In *2011 IEEE third*

international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, 192–199. IEEE, 2011.

Dimitrova, D. V. and Matthes, J. Social media in political campaigning around the world: Theoretical and methodological challenges. 2018.

Federal Election Commission. Federal elections 2016 – election results for the u.s. president, the u.s. senate and the u.s. house of representatives. 2017.

Glaser, J. et al. Auditing political exposure bias: Algorithmic amplification on twitter/x during the 2024 u.s. presidential election. *arXiv preprint arXiv:2411.01852*, 2024.

Hern, A. and Pegg, D. Facebook fined for data breaches in cambridge analytica scandal. *The Guardian*, 2018.

Hugging Face. Hugging face – the ai community building the future. `https://huggingface.co/`, 2024.

Jiang, J., Ren, X. and Ferrara, E. Retweet-bert: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17, 459–469. 2023.

Kim, J., Kim, D. and Park, E. I know your stance! analyzing twitter users' political stance on diverse perspectives. *Journal of Big Data*, Vol. 12(1), 14, 2025.

Marret, C. The impact of social media on elections. *NUPRI-USP Working Papers*, 2020.

Mensah, G. B. Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in ai systems. *Preprint, November*, Vol. 10(1), 2023.

MongoDB, Inc. Mongodb atlas database. `https://www.mongodb.com/products/platform/atlas-database`, 2024.

Neo4j, Inc. *Cypher Manual*, current edn., 2024a.

Neo4j, Inc. Neo4j graph database & analytics. `https://neo4j.com/`, 2024b.

Peng, X., Zhou, Z., Zhang, C. and Xu, K. Online social behavior enhanced detection of political stances in tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18, 1207–1219. 2024.

Reimers, N. and Aarsen, T. sentence-transformers: Embeddings, retrieval, and reranking. `https://pypi.org/project/sentence-transformers/`, 2025. Versión 4.1.0.

Roesslein, J. Tweepy: Twitter for python! `https://www.tweepy.org/`, 2009. Version 4.14.

The Selenium Project. Selenium. https://www.selenium.dev/, 2024.

Trezza, D. To scrape or not to scrape, this is dilemma. the post-api scenario and implications on digital research. *Frontiers in sociology*, Vol. 8, 1145038, 2023.

X Corp. Enterprise api interest form. 2024. Accessed: 2025-04-15.