UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"
PhD SCHOOL

PhD program in: **Public Policy and Administration**

Cycle: **36th**

Disciplinary Field (code): **SPS/04**

A communist, an environmentalist and an android walk into a bar: the measurement and measurable effects of elite communication

Advisor: **Piero STANIG**

Coadvisor: **Nicolò CAVALLI**

PhD thesis by

Henrique MARQUES UCHA MEIRELES ALPALHÃO

ID number: **3111588**

Year 2026

Abstract

There have never been as many things to talk about, or channels through which to talk about them, as there are today. The academic exercise of tracking, collecting and analyzing what is said is an established practice, but its form is in constant and accelerating mutation. In this dissertation, I empirically explore the impact of elite communication on two grand outcomes – national elections and asset pricing – with a strong methodological focus that explores different data collection and treatment procedures and incorporates recent developments in large language model (LLM) technology. I further design and describe an original method for audiovisual data collection that aims to greatly facilitate access to an extremely large and rich, but relatively hard to explore corpus of data, allowing for the extension of existing studies and undertaking of many new ones. By exploring and applying computational tools such as machine learning, LLMs or face and voice recognition, this work stresses the richness of political statement data and strives to demonstrate how to extract the most of and from it.

In the first chapter, through an original dataset of tweets by Portuguese politicians and a natural experiment derived from the outbreak of the Russo-Ukrainian war, I evaluate how stigmatizing behavior towards a radical left party can impact electoral results. Using regression discontinuity design and difference in differences approaches for inference, I obtain results that indicate that the stigmatized party suffers persistent vote intention losses.

The second chapter looks at a Twitter dataset covering German publicly traded firms and Swedish activist Greta Thunberg. It investigates how vocal activism by renowned opinion leaders can create an impact on firms' stock market performance, and how firm behavior might influence this relationship. Results suggest that companies that align themselves with opinion leaders can pass through this process unscathed, while others suffer a stock price decrease. In both the first and second chapter, machine learning and LLM classifiers are employed to refine the data by extracting meaning indicators from the raw text inputs.

The third chapter, finally, describes a method for building an analyzable transcript from audiovisual political data, such as broadcast debates or interviews, allowing for individual speaker diarization and recognition with minimal manual prep-work. This framework can be applied to material in any language, while its agile nature means it is easily adaptable to the specificities of different formats (e.g. short-form social media videos, multi-participant debates, live broadcasts). Through these avenues, it bears the potential to massively expand the amount of available data for political analysis.

This work makes several contributions. Firstly, it employs different methods for the collection and treatment of text, exemplifying their usage, allowing for their comparison, and using them for robust inference. Secondly, it approaches these exercises in a constructive way, aiming to provide better means of obtaining raw data and refining it into its most useful state. It shows, thus, how to employ LLM-based approaches to improve on mainstay methods such as machine learning classifiers or interpolation processes. Thirdly, it introduces a method that provides easily implementable and mostly-automatic access to a particularly rich type of data that was previously hidden behind either notoriously laborious or methodologically complex processes – debate and interview transcripts – and is ready to be adapted to alternative inputs. Insofar as political elites use the different channels at their disposal to convey a cohesive message, these approaches are likely to provide full coverage of politician stances and interventions. However, they go even deeper by tracking each individual to a level of granularity that easily allows for intra-politician message analysis.

Acknowledgements: I would like to thank Marta, who tolerated this; my parents, who made it possible; Piero Stanig, Nicolò Cavalli and José Tavares, respectively my advisors and co-author, who made it better; and Elena, Gian Maria, Paolo and Vini, my cohort, who made it (relatively) enjoyable.

CHAPTER ONE

Parties of the world, ostracize: Political stigmatization of the radical left in a natural experiment setting*

Henrique Alpalhão

Abstract

In an era of remarkable political strife and polarization, parties relate in increasingly acrimonious fashion. Over the last decade, stigma has risen to prominence as a staple of parties' toolkit and a rare source of consensus – across aisles and ideologies, political actors appear to agree that stigmatizing and ostracizing their opponents is a good strategy. Where radicals often target the political establishment and social out-groups, moderates follow suit through campaigns of political discrediting and cordon sanitaire towards radical newcomers themselves. While significant study has already been dedicated to this phenomenon, there are still gaps in coverage, namely in what regards the stigmatization of left-wing parties by the establishment. This work aims to address this specific gap by studying the case of the Portuguese Communist Party (PCP) which, in 2022, became the target of intense and unprecedented stigma by all other parties over their reaction to the outbreak of the war in Ukraine. By using an extremely granular Twitter dataset, I obtain near-universal coverage of party and politician interventions, while the rare natural experiment setting I explore allows for an atypically clean identification of a stigma effect. Using regression discontinuity and difference-in-differences methods, I obtain results that suggest a negative and persistent effect of this new stigma environment in PCP voting intentions.

1 Introduction

> *After decades of hearing PCP criticizing everything and everyone, name-calling and generalizing the other parties, it is priceless to see them bothered by attacks on themselves.*[1]
>
> Duarte Marques (@DuarteMarques), MP (PSD), November 13th 2022

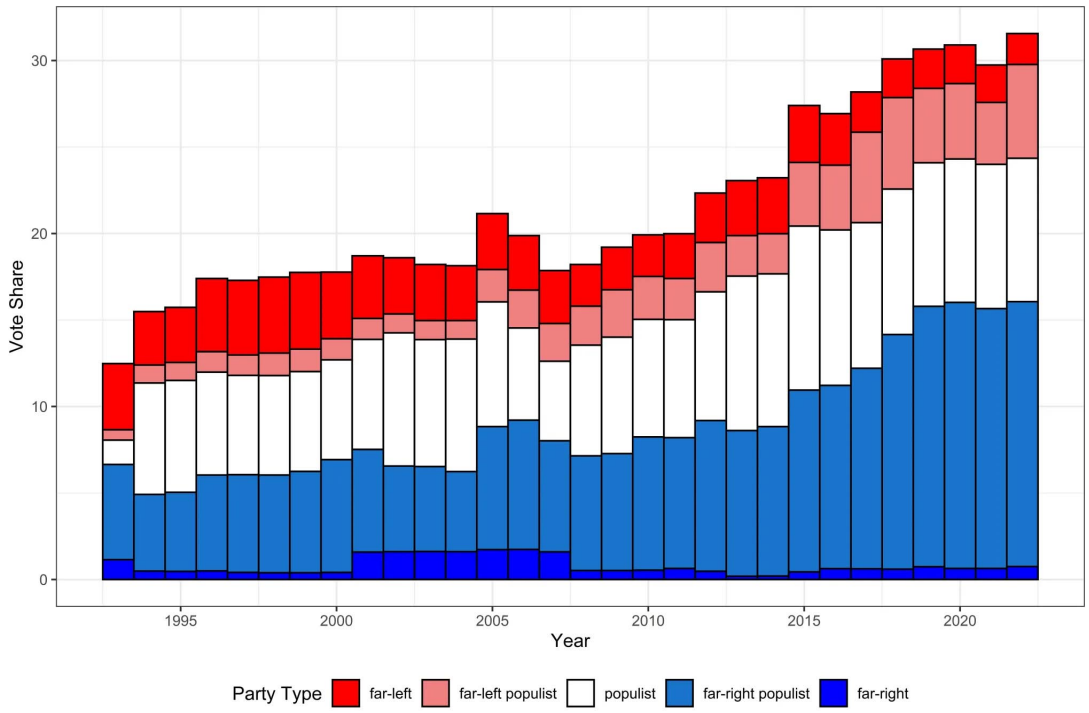
Political radicalism has enjoyed a golden era over the last decade, achieving significant representation gains in European democracies across the world. Figure 1 plots the vote share of populist and radical parties in Europe by year and depicts this trend: from 2010 to 2020, their representation has increased from 20% to over 30%.

These radical newcomers, left- or right-wing, tend to shape political discourse through their own confrontational style and by forcing traditional parties to find strategies that deal with their aggressive posturing (namely by engaging with or ignoring their agenda: see, for example, Meguid, 2005). This generally leads to a more bellicose brand of politics that often descends into ad hominem discourse, the stigmatization of opponents, *cordon sanitaire*

*Comments and suggestions by attendees of the 2024 Elections, Public Opinion and Parties conference regarding this paper are acknowledged and greatly appreciated. This version: November 2025.

[1]Source: https://twitter.com/DuarteMarques/status/1591888580149927938 (in Portuguese).

Figure 1: The PopuList ranking



Notes: Vote shares of populist, far-right and far-left European parties (weighted by population size). Source: The PopuList (https://popu-list.org/).

strategies and an "us vs. them" view of the world. One might even propose a vicious cycle tendency, where the increased presence of this hostile posture breeds more radicalism and risks sucking in or superseding traditional parties.

In such an environment, it becomes relevant to understand the practical (i.e. electoral) implications of the vilification and stigmatization of political opponents, both to inform party strategy and to predict future electoral outcomes. In what follows, I envision political stigmatization as the classification of a given political platform, and by extension their proponents (be they parties or citizens), as irredeemably morally flawed, impossible to converse with, and often incompatible with democratic values. Political ostracism is the practical consequence of this phenomenon – the exclusion of the stigmatized group from political negotiation and consensus-building. These are thus likely to be party-, rather than citizen-driven; arising from inter-politician discourse and party strategies.

Voters might be exposed to this brand of stigma in both their private and public lives – private if this behavior among politicians is mimicked in their interactions with other citizens; and public whenever they witness the occurrence of this phenomenon, even as mere spectators, among politicians – as might occur through broadcast media or exposure to political campaigns. It is, thus, likely to be ubiquitous among at least politically-engaged individuals.

How might voters be influenced by this practice? We can refer to the spiral of silence theory in Noelle-Neumann (1993), through which citizens that hold stigmatized preferences become less likely to express them. If we take a step back, however, to the moment when political preferences are formed, we can also consider the citizen's problem as an allocation one – at each moment considering whether to adopt a stigmatized political preference, if they do not yet hold it; or whether to drop it if they already favor it.

In this case, stigmatization can have a real effect on voting behavior in at least two ways: firstly, if it influences the amount of contact a voter is willing or able to establish with the stigmatized platform. Secondly, if the stigmatization of a party can have an impact in how they perceive it and its valences: a voter that is already partisan

to a stigmatized party might be spurred on to dig their heels in and see their identity reinforced, might instead be disheartened by the lack of popularity of their ideals, or might even be convinced of those ideals' shortcomings. Similarly, a voter who already opposes the stigmatized party might be impacted in both ways; and voters who have not yet decided might begin to envision the stigmatized party as undesirables or, conversely, righteous underdogs. Through these avenues, stigma can and most likely does impact how citizens vote before the ballot – and thus even in secret ballot frameworks.

The literature has dedicated significant attention to the electoral consequences of this phenomenon, with a notable caveat: it looks almost exclusively at radical right-wing parties. This is arguably explained by two factors: firstly, radical right platforms have become increasingly prevalent relative to radical left ones (see figure 1). Secondly, they have also been less successful in achieving social acceptance in western societies than the radical left, and thus might garner more attention from politicians and voters. In any case, a more complete picture can be painted by looking at stigma towards a left-wing platform – if one can be found.

In this paper, I propose an original evaluation of this phenomenon through the specific case of the Portuguese Communist Party (PCP)[2] – an established and electorally relevant player in the system. PCP has had an openly eurosceptic and anti-NATO stance continuously since before the 1974 restoration of democracy in Portugal, has obtained respectable electoral results since elections are held,[3] and generally garnered the respect of the other political parties. Their radical views notwithstanding, they have mostly avoided stigma in the last 50 years. This, however, changed drastically in February 2022, when, in the wake of the Russian invasion of Ukraine, PCP were the only Portuguese party that was perceived as consistently failing to condemn the attack. They have since stuck to their guns,[4] and as such from then on became socially tarnished and the target of unprecedented stigma by other politicians.

This is a relevant enterprise for several reasons. Firstly, it provides a cleaner natural experiment identification of a phenomenon that is extremely hard to isolate (does stigma cause or track electoral results?). Secondly, it documents an instance of stigma against an established radical left-wing party, rather than recent radical right parties. Thirdly, unlike a significant part of the literature, it studies politicians, rather than voters. Finally, drawing on polls and Twitter data, it uses an extremely granular dataset that arguably covers most of the universe of public politician interventions.

2 Literature review

> *Protesters being murdered in cold blood. PCP defending the murderers. PCP has been doing this for years almost without any social backlash. To the left, there is complete silence. This is not normalizing barbarism, it is being complacent with it.* [5]
>
> Carlos Guimarães Pinto (@carlosgpinto), MP (IL), August 16th 2020

This section discusses the relevant literature for the paper, which can be divided into three main topics: stigma, communist parties and the communist vote, and the use of Twitter to study political phenomena.

2.1 Stigma and ostracism in modern politics

Political stigma can be defined as the perception that given parties and their voters hold unacceptable views and are politically illegitimate. It can reveal itself at the party and voter level, and might aim to shame a given sector of

[2] In Portuguese: *Partido Comunista Português*. Appendix A1 briefly introduces them and all other mentioned parties.

[3] Figure 2 in section 3 presents this data.

[4] See, for example, the following article by Novaya Gazeta Europe, which finds that PCP MEPs are those that most abstain or vote against anti-Russia resolutions: https://novayagazeta.eu/articles/2023/02/27/putins-sidekicks-en.

[5] Source: https://twitter.com/carlosgpinto/status/1294931284482052097 (in Portuguese).

the electorate or party system into reclusion or conformity. As Valentim (2021) discusses, there is ample evidence that social norms constrain political behavior towards conforming to the norm. This is aligned with Kuran's (1987) interpretation of the stigmatized citizen's problem: a tradeoff between saving their social reputation or sticking to their true preferences.[6]

Relevant work has been made on stigma and resulting "shame effects" – the phenomenon by which voters feel compelled to hide their true preferences. Firstly, Valentim (2021) finds that achieving parliamentary representation is the most relevant factor in normalizing radical right support and reducing shame among radical right voters, the mechanism potentially being that representation might act as a societal signal that supporting the new party is acceptable. This fits into the Bischof and Wagner (2019) suggestion that the parliamentary entry of a radical right party exacerbates polarization through two avenues: the legitimization of more extreme proposals and backlash from their more established peers. Valentim (2022) further confirms that individuals with stigmatized preferences have an incentive to conceal them by exploring a design flaw in Spanish local elections that, in some municipalities, meant the vote was not secret. Finally, Valentim and Widmann (2023), in an analysis of German parliament speeches, find ample evidence of radical right shaming by politicians of other political inclinations, showing that this phenomenon occurs between politicians as well as voters.

On the closely related topic of inter-party political ostracism, Joost van Spanje and coauthors provide significant contributions. Van Spanje and Van der Brug (2007) focus on the radical right and define ostracism as a situation where other political parties reject cooperation with a given party (e.g. forming coalitions, joint lawmaking) on the basis of principle, due to their perception as anti-democratic. They argue that ostracism, for the mainstream parties, serves purposes such as the prevention of power sharing and vote "wasting". The delegitimization of far-right parties is meant to keep them from gaining representation and to keep voters from diverting right-wing votes to the more extreme option (Art, 2007). They further find that their inclusion in the system by mainstream parties (e.g. FPÖ in Austria) tends to moderate the action of extreme parties as they require compromise to achieve meaningful representation, while ostracism creates no such incentives and leads to sustained extremism. Additionally, ostracism may exacerbate existing feelings of exclusion from democratic rights within far-right groups, leading to stronger bonding between party supporters and potentially more extreme positions than before.

Van Spanje (2010), on the other hand, provides an analysis of the factors that determine whether a far-right party is politically shunned or not. He finds that the main drivers are size and ideology: the smaller the extremist party, the lower its political usefulness for larger parties (and thus the higher the likelihood of ostracism is), and the further away ideologically the two parties are, the more likely ostracism is to occur.

On the impact of this phenomenon on voting outcomes, Han (2020) concludes that exclusion leads to increased far-right support by voters with authoritarian inclinations, but decreases it otherwise – again exacerbating the extreme positions that are usually associated with these parties.

While the above-mentioned studies appear to paint a coherent picture, the literature does not all agree on the workings and effects of ostracism – notably, Akkerman and Rooduijn (2015) find that, at least regarding immigration and European integration, non-ostracized parties have become just as radical as those who were ostracized. This phenomenon, thus, still warrants additional study.

2.2 Twitter as a stage and data source

The literature offers a reasonable corpus of considerations, both theoretical and practical, on the nature and usage of Twitter data for social sciences. Firstly, Barberá and Steinert-Threlkeld (2019) provide a useful theoretical discussion of social media use for political science research. While they note that the data-generating process is not necessarily known (e.g. the platform could have been running an experiment or blocking certain content during data generation and collection), they nonetheless identify several relevant advantages brought by this kind of data.

[6] See also the spiral of silence in Noelle-Neumann (1993).

Some examples are its unparalleled granularity; the lower likelihood of behavior changes from observation (as agents are being observed in a real-world environment rather than an experiment); the finding that offline behavior does mimic online conduct; and the achievability of representativeness for elite studies.[7] The latter point is especially relevant for this work, which uses politician rather than voter tweets. Barberá and Rivero (2015) further provide a good literature review on the use of Twitter data for political science research. They also reinforce that Twitter data is generally not representative if the general public is being studied: extreme positions and urban topics are overrepresented as the majority of tweets come from a minority of "hyper-active" users.

Another strand of this literature offers more practical considerations. Lietz et al. (2014), for example, using a dataset of 2013 tweets by German politicians and the public, find that message volumes increase around election times. Theocharis et al. (2020), in a slightly different vein, provide a descriptive account of uncivil interactions between citizens and politicians on Twitter. They find that a sizable 18% of all tweets mentioning US Members of Congress in their sample are uncivil, and that politicians themselves tend to not engage with their audiences via these channels, limiting their participation to broadcasting purposes. Finally, Ernst et al. (2017) perform content analysis on party social media communication in six western democracies and find that populism manifests itself on both sides of the political extremes. They suggest that social media is especially useful for these movements as they provide cheap and unmediated access to voters and circumvent some gatekeepers (e.g. media outlets).

2.3 The modern communist vote

Finally, and before discussing the specific case of PCP, it is relevant to establish some electoral and ideological context for the communist parties of Western Europe. To do so, I begin by drawing on Sassoon (1992), who delves into the post-WW2 period. He discusses how pro-capitalist platforms lost electoral support across polities, leading to gains for left-wing parties and a reframing of the right towards more social concerns (as in Christian democracy, for example). Of the former, communists were generally credited with being more staunchly opposed to fascism than socialists, and correspondingly saw larger rewards in terms of votes.[8] Sasson (1992) presents communist vote shares in the first post-war elections in several West European countries (p. 152), where some remarkable results can be seen (e.g. 26% in France, 23.5% in Finland, 19% in Italy). As I will discuss in the next section, this phenomenon is closely related to what happened in Iberia when, 30 years later, communist movements were integral parts of the resistance to their authoritarian governments.

After this initial boom, however, the communist vote began to decline fairly quickly. Starobin (1965) discusses the then-current moment of communism in the region. He states that "communism is a factor, sometimes serious, sometimes vestigial in Western Europe today (...)." A few causes for this are identified in the literature. Firstly, Starobin argues that communist parties were hurt by their lack of "European solidarity", establishing connections and cooperation almost exclusively with the Soviet Union and behaving as outposts that were to hold and prepare for a political shift that could favor the USSR. Sassoon (1992) further adds that the advent of the Cold War made other parties extremely reluctant to discuss with and include communists; and that the perceived main difference between communists and socialists was the former's stronger militancy – a characteristic electorates had come to distrust following the Second World War.

In general, Sassoon (1992) states that communism retained relevance only where it managed to surpass social democracy to become the main representative of the "left radical tradition", be it democratically (e.g. France, Italy) or through ongoing political resistance, as in Spain and Portugal.

Lazar (1988) provides a good account of the situation by the 1980s, characterized by a general electoral decline of communism in Europe. He singles out the Cypriot and Italian communist parties as the then-most powerful,

[7] Barberá and Rivero (2015): "virtually, all candidates and elected officials have a presence on Twitter." For examples of elite studies, see Barberá et al. (2019) and Fazekas et al. (2021).

[8] Sassoon (1992): "There is some correlation between the magnitude of the resistance and communist political successes immediately after the war."

and their Portuguese and Greek counterparts as relevant, relative newcomers, having emerged from clandestinity in 1974 in both countries. The author provides membership figures[9] for parties in the region that, while hard to confirm, could provide interesting insight: in 1984, the French Communist Party boasted a membership 380 000 strong (140 000 less than in 1978); in 1986, the Italian Communist Party claimed to have 1 596 000 members (a 218 000 drop since 1976); that same year, their Spanish counterpart had 60 000 (110 000 less than in 1978). Against this backdrop, the Portuguese Communist party claimed to be experiencing a massive boom in membership: from 15 000 in 1974 to 201 000 in 1983.

In general, the author observes that communist parties were becoming less popular among young and working class people as unionism declined, and that even those parties that were influential in their political systems had to settle for a position of dependence in regards to other parties. He adds that PCP and their Greek and Cypriot counterparts were the only West European communist parties that had managed to avoid social backlash against their activity – a topic I will further touch upon in the following sections.

3 Communism and the Portuguese Communist Party

> *100 years of PCP. Happy birthday! Democracy owes them an immense sacrifice in the long night of fascism. Yes, it's true. Portugal was once silenced.* [10]
>
> Eurico Brilhante Dias (@EBrilhanteDias), MP (PS), March 6th 2021

> *The Portuguese Communist Party has never stopped believing in the power of the state over the individual and in the capacity of an enlightened elite to decide at every moment what is best for each citizen. It has been a century of consistency in the defense of dictatorships, worship of inhumane leaders, and persistent support for regimes that spread hunger, misery, and death.*
>
> *Iniciativa Liberal wants to take this opportunity to congratulate the Portuguese people for never allowing the Portuguese Communist Party to destroy democracy and freedom in our country. (...)* [11]
>
> Iniciativa Liberal (@LiberalPT), official party account, March 7th 2020

The Portuguese Communist Party (PCP) was founded in 1921, during the tumultuous First Republic, and is the oldest active Portuguese party. This section provides ideological and electoral context for PCP since its inception, based mainly on Lisi (2008) and Patrício and Stoleroff (1994).

3.1 Electoral dynamics, 1974-2023

Lisi (2008) discusses PCP and its evolution in the revolutionary and transition to democracy period.[12] The April 1974 coup, led by a military movement called MFA (*Movimento das Forças Armadas* – "Armed Forces Movement"), overthrew the *Estado Novo* regime ("New State", 1933-1974), one of the longest-lived dictatorships in 20th-century Europe. Immediately after the coup, political power was concentrated in a military junta,[13] with the provisional government, composed by the leaders of the main political parties, meant to simply implement the designations of

[9] Lazar (1988), pp. 245-246.

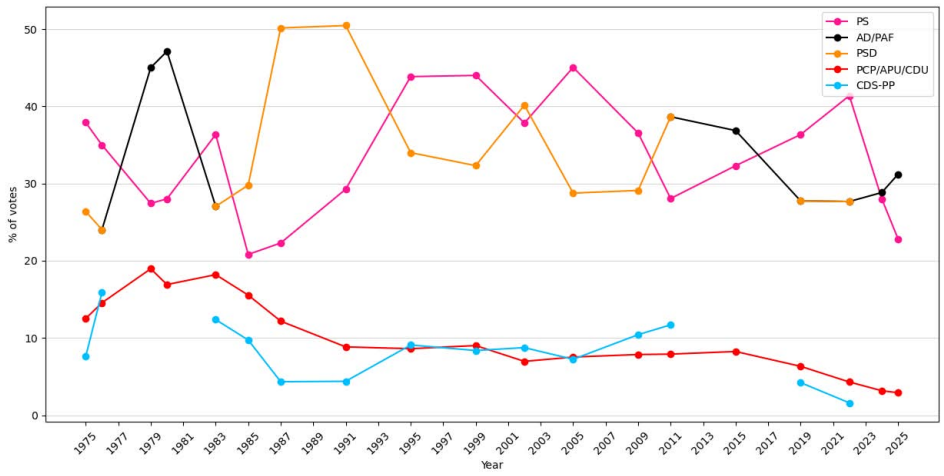
[10] Source: https://twitter.com/EBrilhanteDias/status/1368211646183841797 (in Portuguese).

[11] Source: https://twitter.com/LiberalPT/status/1236418812943138817 and https://twitter.com/LiberalPT/status/1236418814629339136 (in Portuguese).

[12] For context on the party's clandestine, pre-1974 period, refer to Madeira (2013). For alternatives in English, see Cunha (2020) or the official English-language PCP program at https://www.pcp.pt/sites/default/files/documentos/pcp_programme_and_constitution_approved_by_the_19th_congress.pdf.

[13] *Junta de Salvação Nacional* – "National Salvation Junta".

Figure 2: Legislative election results for selected parties, 1975-2025



Notes: This graph compares selected founding parties of the Portuguese Third Republic. Author's own construction with data from Marktest (https://www.marktest.com/wap/a/p/id~13d/el~0.aspx).

the junta and MFA. In practice, MFA ruled without democratic restrictions through "revolutionary legitimacy". As Lisi discusses, PCP became intertwined with MFA,[14] was significantly ahead of all other parties organizationally (as the others had just been founded), and enjoyed some "backward legitimacy" from its resistance activities in the clandestine period.

These institutions contained several different factions (PCP, specifically, advocated for a military-backed, USSR-inspired model) whose clashes occasionally threatened to escalate into armed conflict until the first legislative elections in 1976. These elections yielded a parliamentary majority for the Socialist Party (PS), dictating the victory of the moderates and the introduction of a liberal democracy. On this topic, the 1975 interview granted by Álvaro Cunhal, then PCP's leader, to the Italian journalist Oriana Fallaci left no doubts as to the party's position, as the following excerpts show: "But we, the communists, do not accept the game of elections. (...) No, no, no, elections do not interest me at all. Nothing at all!"; "I assure you that in Portugal there will be no Parliament."; "Portugal will not be a country with democratic freedoms (...). It will not be a companion to your bourgeois democracies. Because we will not allow it."[15]

Figure 2 shows voting trends for electorally relevant parties in 1975. Note that PCP has not run for elections by itself since 1976 – rather, they since led coalitions of like-minded parties under the APU (*Aliança Povo Unido* – "United People Alliance") and CDU (*Coligação Democrática Unitária* – "Democratic and Unitary Coalition") names. In what follows, I refer to PCP by the appropriate coalition name when discussing election results. Under these designations, the figure shows how the communists obtained a vote of above 10% until 1991 (even if with a downward trend from 1983 onward) and that, unlike the other parties, their vote seems to be relatively unaffected by electoral cycles.

Patrício and Stoleroff (1994) discuss the party's situation and strategy in the 1980s, specifically in light of the Perestroika era. They describe PCP as, until very recently, "a bastion of communist orthodoxy modeled along Soviet lines (...) [which] still manifested characteristics that could fairly be labelled 'Stalinist'". In face of very

[14] A relationship that would become much weaker by the end of 1975, in MFA's final months.

[15] Fallaci (1975). Original excerpts (in Portuguese): "Mas nós, os comunistas, não aceitamos o jogo das eleições. Você equivoca-se ao partir desse conceito. Não, não, não a mim não interessam nada as eleições. Nada mesmo!"; "Asseguro-lhe que em Portugal não haverá Parlamento."; "Portugal não será um país com liberdades democráticas e monopólios. Não será um companheiro de viagem das suas democracias burguesas. Porque não o permitiremos." A copy can be found at https://hemerotecadigital.cm-lisboa.pt/Periodicos/jornaldocasorepublica/N08/N08_master/JornaldoCasoRepublica_N08.pdf.

significant changes such as Glasnost, Perestroika and Portuguese accession to the EEC, the party persisted in its orthodoxy, leading to significant dissent among their lines and to several members being expelled or abandoning the party of their own volition. Patrício and Stoleroff (1994) also present numbers that show that PCP membership decreased between 1983 and 1988, in contrast to the sharp increase in the 1974-83 period described in section 2.3. The authors describe this era as one of resistance to change by PCP, which led to defections,[16] loss of popular support, the aging of their members, and ultimately a loss of relevance and move from an active to a reactive role in Portuguese politics.

Between this point and 2015, PCP remained reasonably immobile, both ideologically and electorally (at around the 8% mark), effectively relegated to a role of perpetual opposition as PS and PSD alternated in power. 2015, however, brought about a change in paradigm. Following that year's elections, and despite having been the second-most voted party, PS managed to form a government through the support of a post-electoral left-wing coalition – namely[17] BE, PCP, PEV, and PAN. This unit, nicknamed *Geringonça* ("contraption"), allowed PS to effectively govern without a majority and to see its tenure to its conclusion in 2019. Although this government was composed entirely of PS elements, the radical left was brought to a prominence that had not been seen for more than two decades.

PS attempted to replicate this solution after winning the 2019 elections without a majority, with one relevant difference – the lack of any formal agreement, instead relying on piecemeal negotiation with their partners. This proved less robust when, in October 2021, BE, PCP and PEV joined the right-wing parties in voting against the government's 2022 general budget proposal. Given the socialists' apparent inability to continue governing effectively, parliament was dissolved and new elections scheduled for January 2022.

These elections ushered in significant changes. PS were those who reaped the most rewards, taking a significant share of their previous partners' vote and attaining an absolute majority. The new right also benefitted: both *Chega!* and IL had their best result to date, electing 12 and 8 MPs to become the third and fourth most represented parties, respectively. CDU, on the other hand, suffered quite significantly by obtaining their worst result ever (4.39%) and losing 6 MPs (specifically, four PCP and two PEV, dictating the latter's loss of parliamentary representation). In 2024 and 2025, two new snap elections solidified the first right-wing majority in more than a decade and brought further bad news for PCP: two new all-time low results for CDU (3.17% and 2.91%, respectively), and two corresponding drops in the number of PCP MPs (from six in 2022 to four in 2024 and three and 2025). These parliamentary configurations are depicted in Figure 3.

3.2 Position regarding the Russo-Ukrainian war

Most of PCP's ideological stances, as the previous sections have shown, did not change much since before the dissolution of the USSR. This is the case regarding their view of NATO and other Western international organizations (e.g. the EU), as described by Patrício and Stoleroff (1994), Lisi (2008), and current-day political interventions. We can use PCP tweets to illustrate how their stance in this regard was not altered by the eruption of the war:

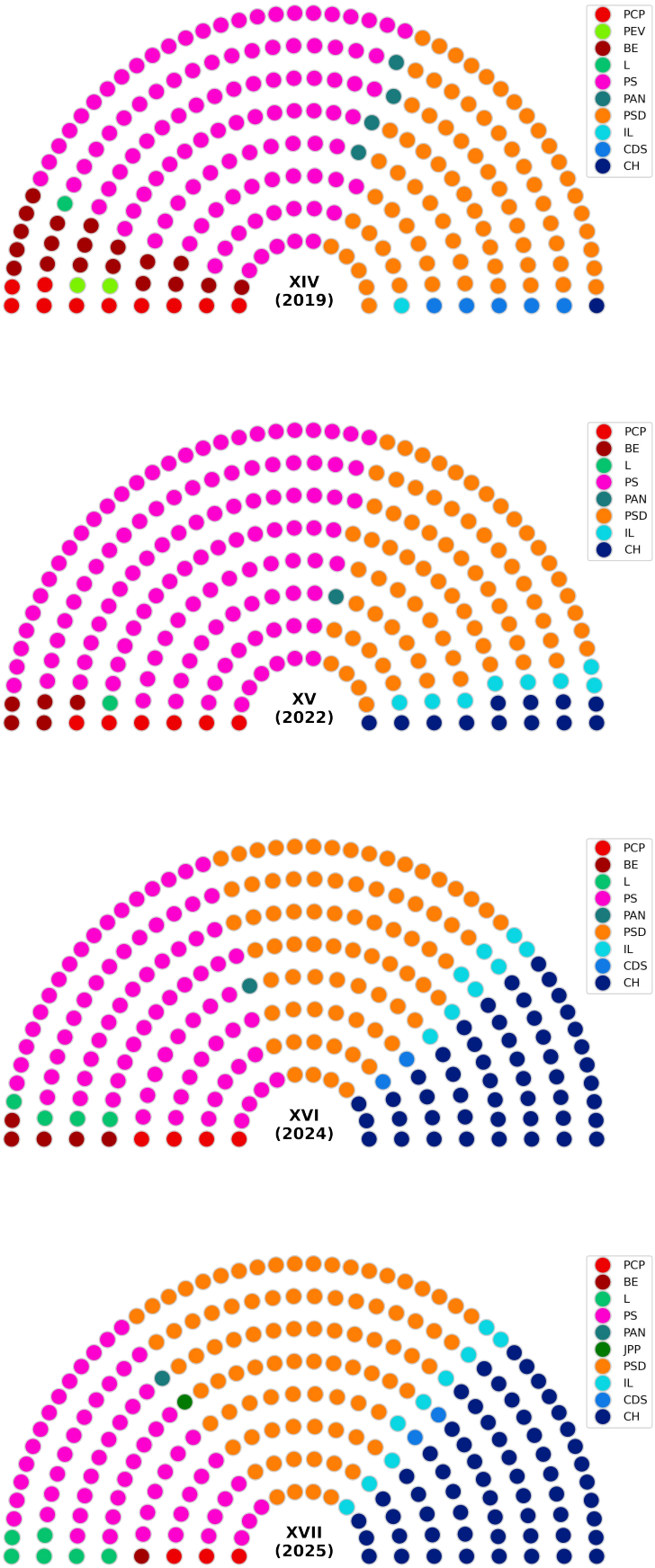
> *On April 4th, 1949, NATO was founded, of which Portugal, under fascist dictatorship, is a founding member. Throughout 71 years, it has been one of the most aggressive and deadly instruments of imperialism, as the peoples targeted by its aggressions do not forget: NATO means war.* [18]
>
> PCP official account (@pcp_pt), April 4th 2020

[16] Several of these defectors would go on to have relevant careers representing other parties, such as Zita Seabra (PSD and, more recently, IL), Vital Moreira (PS), and José Magalhães (PS).

[17] And in order of parliamentary representation.

[18] Source: https://twitter.com/pcp_pt/status/1246512864770179072 (in Portuguese).

Figure 3: Portuguese Parliament, 2019-2025



(...) • Actions for peace and against war continue

• USA, NATO, and EU meet in Brussels to intensify confrontation (...)[19]

PCP official account (@pcp_pt), March 24th 2022

PCP's stance on Ukrainian politics is also found to have been unchanged and openly stated since before the war, as these tweets, from 2020 and 2022, respectively, show:

The script of the bloody coup of Maidan in Kiev remains omnipresent, but Belarus is not Ukraine, where neo-Nazi militias march on the streets today under the aura of power. [20]

PCP official account (@pcp_pt), August 31st 2020

July/2014: PCP defends peace in Ukraine, shows solidarity with the Ukrainian people, denounces the instrumentalization by the USA/EU/NATO, and the actions of the far-right. It also condemns political persecution and military attacks against the people. We didn't just remember Ukraine on February 24, 2022. [21]

João Oliveira (@joao_g_oliveira), former MP (PCP), March 5th 2022

These stances were notoriously crystalized in April 21st 2022, when Volodymyr Zelenskyy, the Ukrainian president, spoke to Portuguese Parliament.[22] PCP MPs boycotted this session, drawing comparisons to May 9th 1985, when they had done the same in protest of another foreign dignitary's formal reception in Parliament – then-US president Ronald Reagan.[23]

4 Data and framework

Might the Ukraine war not be a television recreation of Orson Welles' War of the Worlds? [24]

António Filipe (@AntonioFilipe), former MP (PCP), February 15th 2022

This work's goal is to evaluate the impact of stigmatization on the stigmatized party's voting intentions in circumstances that allow for an unusual isolation of the stigma phenomenon. Why is identification an issue? In looking for the impact of stigma on voting intentions, one must bear in mind that stigma is usually a response to a behavior that is perceived as negative. As such, the cause for stigma itself might have an impact on voting intentions, meaning a more straightforward operationalization would risk suffering from endogeneity. There is also scope for reverse causality in this relationship – while increased stigma is likely to impact vote intentions, it is also plausible that parties that poll better face increased stigma from their opponents as an electoral strategy.

To tackle this issue, I study the specific case of PCP and anti-PCP stigma by the other Portuguese political parties and their main figures. In this section, I discuss the quantification of stigma and vote intentions, make a case for the natural experiment I propose to explore, and devise an estimation framework based on regression discontinuity and difference-in-differences methods.

[19]Source: unavailable (most likely deleted). Original (in Portuguese): "• Comunistas reclamam viragem nas políticas ambientais • 11.ª Assembleia da OR de Aveiro do PCP: «Organizar, Lutar, Avançar» • Prosseguem as acções pela paz, contra a guerra • EUA, NATO e UE reúnem-se em Bruxelas para intensificar confrontação. Tudo em https://t.co/5NuplbGg3C."

[20]Source: https://twitter.com/pcp_pt/status/1300381064175775744 (in Portuguese).

[21]Source: https://twitter.com/joao_g_oliveira/status/1500068855410003969 (in Portuguese).

[22]Footage of which can be found at https://canal.parlamento.pt/?cid=5855&title=reuniao-plenaria.

[23]Footage can be found here: https://www.youtube.com/watch?v=sYm_knZWt9Q.

[24]Source: https://twitter.com/AntonioFilipe/status/1493414338371231751 (in Portuguese).

The main premise of my estimation strategy is simple: PCP faced an unprecedented and persistent wave of stigmatization, namely by their politician peers, immediately after and as a result of the Russian invasion of Ukraine. Given the ideological and policy platforms that justify their position towards the conflict have been solidified and publicized for decades, as the previous sections show, one may plausibly argue that the only new phenomenon that could have impacted vote intentions at this time was this stigmatization – the invasion of Ukraine brought about unprecedented and unpredicted stigma, but no new information on party stances that might inform vote intentions, allowing for the isolation of a stigma effect.

4.1 Stigma and its identification

My formulation of stigma – perhaps best described as inter-party stigma – implies a generalized stance among political parties that singles out a given party as an untenable out-group that does not belong in the system and with whom one should not cooperate. When this phenomenon occurs, politicians regularly assert these positions in their public appearances (e.g. parliamentary speeches, interviews, rallies). For the purpose of this paper, I use the Twitter presence of Portuguese politicians to identify the intensity of anti-PCP stigma, with the two aforementioned advantages of near-universal coverage (politicians who shape public discussion are active on Twitter) and accurate representation of their overall messages (what politicians say on Twitter is representative of what they say through other means of communication, e.g. Casas and Morar, 2015). Two examples, of a stigmatizing and non-stigmatizing tweet about PCP, respectively, can be found below:

> *Great caricature of a despicable bootlicking, attached to a defamatory campaign that violates criminal law and turns the Kremlin's spokespersons into common criminals. They tarnish the name of the institution to which they belong and venerate a horrendous war criminal.* [25]
>
> José Magalhães (@zmaglh), MP (PS), April 23rd 2022

> *This might be news to some, but there is no far left in Portugal. There used to be, like the Revolutionary Brigades (PRB) or the Global Project (which included the FP-25). Fortunately, the Portuguese Communist Party (PCP) resisted those who veered towards such extremes and continues to play a crucial role in democracy to this day.* [26]
>
> Isabel Moreira (@IsabelLMMoreira), MP (PS), September 26th 2022

To quantify this phenomenon, I begin by gathering a database of all tweets by Portuguese politicians in the January 2010-May 2023 timeframe. This adds up to 614 669 tweets, 180 politician accounts and one aggregator account that allows me to recover some tweets that were deleted by their authors.[27] Appendix A3 lists all these accounts, as well as their total tweet output, role and party affiliation.

I then train three classifiers that label each tweet as mentioning PCP (regardless of the message) or not. The first two are supervised bag-of-words classifiers trained with logistic regression, and as such require a training set. To obtain it, I manually label a sample of 12 269 tweets that included the following frequently PCP-related words:

[25] Source: https://twitter.com/zmaglh/status/1517806816540954624 (in Portuguese).
[26] Source: https://twitter.com/IsabelLMMoreira/status/1574439809485078529 (in Portuguese).
[27] The impact of these is negligible, as they account for only about 3500 tweets (likely much less than the actual number of deleted tweets by these accounts), but I include them nonetheless for completeness. Deleted tweets were recoverable only until May 2022, when the aggregator went silent. Accounts for which at least one deleted tweet was recovered are as follows: PCP (@pcp_pt), CDS-PP (@_CDSPP), Livre (@LIVREpt), Os Verdes (@OsVerdes), PSD (@ppdpsd), PS (@psocialista), António Prôa (@antonioproa), Edite Estrela (@editeestrela), João Galamba (@Joaogalamba), Rui Tavares (@ruitavares), António Filipe (@AntonioFilipe), Assunção Cristas (@CristasAssuncao), José Magalhães (@zmaglh) and Carlos Moedas (@Moedas).

"PCP", "PZP",[28] "USSR", "soviet", "comuna",[29] "far-left", "communism", "communist", "Lenin", "Stalin", "Venezuela", "North Korea", "dictatorship", "China", "avante",[30] "comrade", "Jerónimo",[31] "Raimundo."[32] By employing both an all features (i.e. consider all terms that might correlate with the classification) and a selected features (consider only the terms that correlate the most) approach, this yields two alternative classifiers.

The final classifier is a GPT-based zero-shot task that, for each tweet, passes the following prompt to GPT-4.1:

> *You are a research assistant classifying tweets by Portuguese political parties.*
>
> *Decide whether the tweet refers to the communist CDU coalition or to one of its member parties (PCP or PEV) in ANY way.*
>
> *Tweets about CDU might display or mention one or more of the following:*
>
> • *Explicit party names or abbreviations: CDU, "Coligação Democrática Unitária", PCP, "Partido Comunista Português", PEV, "Os Verdes"*
>
> • *Party Twitter handles (@pcp_pt, @CDUPCPPEV, @OsVerdes)*
>
> • *Current or past political figures: Jerónimo de Sousa, Paulo Raimundo, João Ferreira, João Oliveira, António Filipe, Paula Santos, Álvaro Cunhal, Heloísa Apolónia*
>
> • *Generic mentions of communism, historical or current communist states (e.g. USSR, Cuba, Venezuela, North Korea), and communist leaders (e.g. Stalin, Lenin, Fidel Castro, Hugo Chávez, Kim Jong-un)*
>
> *Label each tweet with:*
>
> *1 – the tweet references CDU or PCP/PEV (directly or indirectly)*
>
> *0 – no reference, or you are unsure*
>
> *Please begin your answer with your chosen label (1 or 0), followed by a short explanation.*
>
> *Tweet: {text}*

Each of these classifiers has particular strengths and weaknesses that mean that they can be complementary: the all features classifier, by considering all terms that correlate with mentioning PCP, can identify positives through relatively rare avenues, but might also mislabel when encountering broader terms – a frequent cause of this issue is "far-left", which occurs often when discussing the communists but can also appear in reference to other parties. The selected features classifier is less likely to suffer from the latter issue, but can also lose reach by limiting the amount of ways to identify whether a tweet refers to PCP or not. The GPT classifier, finally, is uniquely able to identify nuance and tacit references, but often overreaches by making connections that are too far-fetched (as a result, when applied to the entire dataset it finds many more positives than the other classifiers: 8452, as opposed to 4356 and 4242 for the all and selected features, respectively).

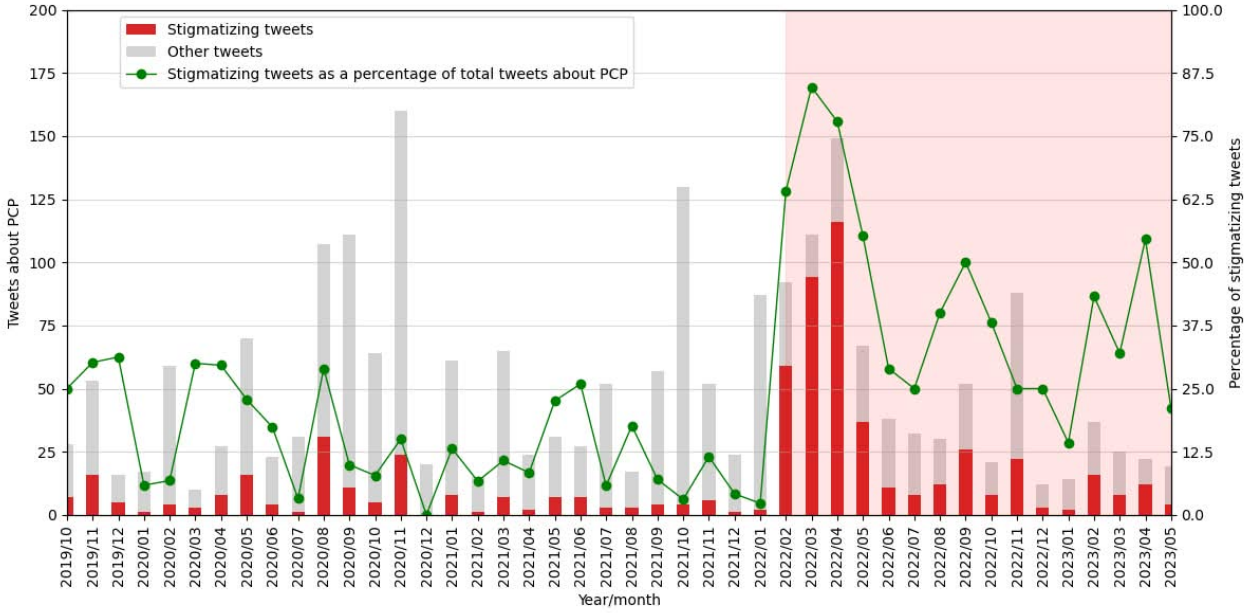
To take all of these into account, I implement an ensemble classifier, which considers a tweet as being about PCP (Ab_PCP=1) if at least two of the three aforementioned classifiers do so.

The Ab_PCP classification task, after dropping all tweets by PCP, PEV and CDU, yields a total of 4 306 tweets about PCP, which rise to 4 511 if enforcing the manual labels previously used for training. Given the manageability of this figure, I manually classify each Ab_PCP=1 tweet as being stigmatizing or non-stigmatizing towards the party. In general, a stigmatizing classification implied referring to PCP as behaving in a shameful or unacceptable way, comparing it with the USSR or other dictatorships, or statements of anti-communism.

Table 1 presents stats and accuracy metrics for these tasks, showing that, although all classifiers perform well,

[28] A common formulation in 2022 that associates PCP with Russia's "Z" war symbol.
[29] Meaning "commie."
[30] A catchphrase among communists (meaning "ahead" and implying movement towards victory) and the name of PCP's official newspaper.
[31] Referring to Jerónimo de Sousa, PCP's Secretary General until 2023.
[32] Referring to Paulo Raimundo, PCP's Secretary General since 2023.

Figure 4: Stigmatizing and non-stigmatizing tweets about PCP per month



Notes: CDU, PCP and PEV tweets are excluded. The area shaded in red represents the war in Ukraine.

the ensemble formulation performs best.[33] Figure 4, on the other hand, shows total and stigmatizing tweets per month across time (a view of the entire timeframe can be found in figure A1, in the appendix), while figure 5 compares these two series per party in the 15 months before and 15 months after the war broke out.

From figure 4, three features are of particular interest. Firstly, a very sizable jump in stigma can be seen from January to February 2022, both in the absolute number of stigmatizing tweets, from under 10 to around 60, and in their percentage relative to total tweets, from close to zero to nearly 65%, a figure to rises to over 75% in March and April. Secondly, the war seems to have brought about a change in paradigm, at least in the timeframe under analysis – overall, the amount of non-stigmatizing tweets consistently dropped while the number of stigmatizing ones consistently rose, leading, on average, to a higher representativity of stigmatizing messages per month when compared to the pre-war period. Thirdly, stigma is found to be particularly intense in the February-May 2022 interval, indicating that the effect of the war on stigma, though apparently persistent, was particularly strong in the few months after the war broke out.

Figure 5 allows for further relevant insight. Firstly, it clearly shows how all parties with parliamentary representation significantly increased their stigmatization of PCP after the war broke out. It is further noteworthy how only *Chega!* began tweeting significantly more about PCP in general after the war broke out (possibly to politically capitalize on the topic, as *Chega!* and PCP directly dispute some electoral districts) – PSD, PS and IL's total mentions remained relatively constant, while CDS, BE, L and PAN's decreased. It is especially relevant to see that BE were among those who most refrained from discussing PCP, in a stigmatizing way or otherwise – as I will discuss in the next section, this is likely due to their initially inconsistent position regarding the conflict.

Together, these figures make a strong case for discontinuity – both in aggregate terms and for each party, the war dramatically increased the amount of stigma hurled at PCP and decreased other, non-stigmatizing mentions. It is also noteworthy that parties were already tweeting about PCP before the outbreak of the war, allowing me to more credibly ascribe the change to stigma, rather than message volume.

[33]Note that the classifier does not perform perfectly on the manually-labelled tweets – in the few cases when they are misclassified, the manual labels are enforced. This means actual accuracy metrics will be slightly better than those presented.

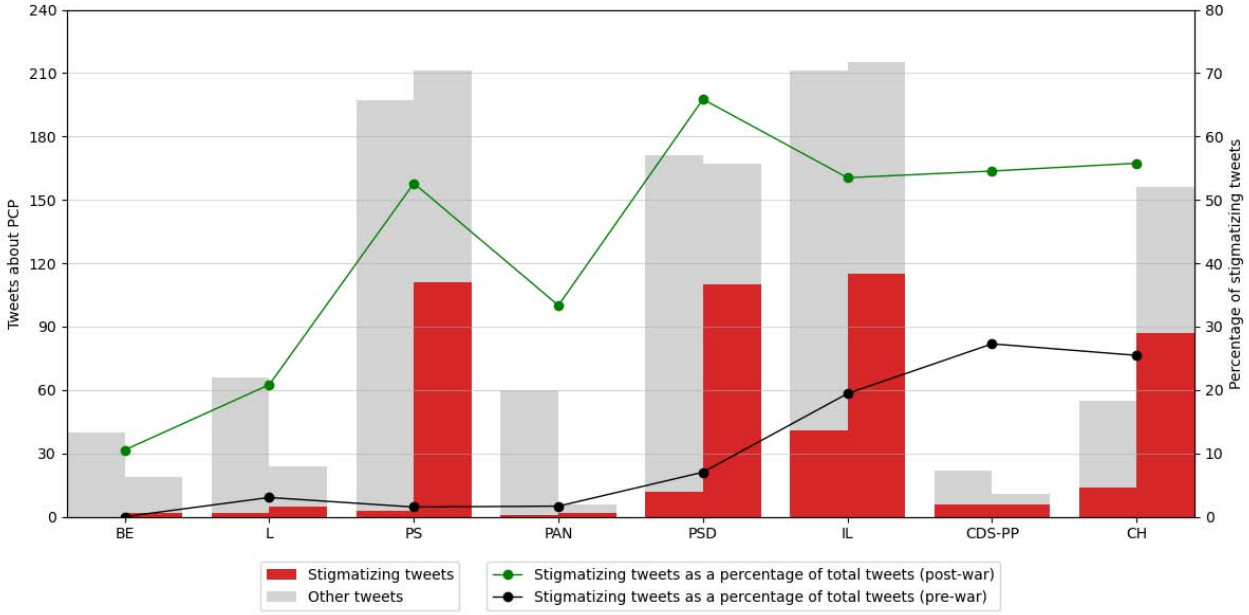
Table 1: Classifier metrics

All features				
	Precision	Recall	F1	Support
0	0.95	0.98	0.96	1883
1	0.91	0.82	0.87	571
Accuracy			0.94	2454
Macro Avg.	0.93	0.99	0.91	2454
Weighted Avg.	0.94	0.94	0.94	2454
Selected features				
	Precision	Recall	F1	Support
0	0.94	0.98	0.96	1883
1	0.92	0.80	0.86	571
Accuracy			0.94	2454
Macro Avg.	0.93	0.89	0.91	2454
Weighted Avg.	0.94	0.94	0.94	2454
GPT				
	Precision	Recall	F1	Support
0	0.97	0.93	0.95	1883
1	0.81	0.91	0.85	571
Accuracy			0.93	2454
Macro Avg.	0.89	0.92	0.90	2454
Weighted Avg.	0.93	0.93	0.93	2454
Ensemble				
	Precision	Recall	F1	Support
0	0.95	0.98	0.96	1883
1	0.92	0.83	0.87	571
Accuracy			0.94	2454
Macro Avg.	0.94	0.90	0.92	2454
Weighted Avg.	0.94	0.94	0.94	2454

	=0	=1
Ab_PCP: Training set	9 412	2 857
Ab_PCP: Full dataset (ENS)	580 117	4 306
Ab_PCP: Full dataset (ENS, coerced)	579 912	4 511
Stigma (manual)	3754	757

Notes: all figures refer to the entire tweet dataset excluding tweets by PCP, PEV or CDU accounts.

Figure 5: Stigmatizing and non-stigmatizing tweets about PCP by party, before and after the war



Notes: This plot depicts the tweeting behavior of each party in the 15 months before (left, November 2020 - January 2022) and 15 months after (right, February 2022 - May 2023) the outbreak of the war.

4.2 Vote intentions

To measure voting intentions, on the other hand, I use a database of all polls and legislative election results since 2010, which, as of September 5th 2025, comprises 656 polls corresponding to 578 different days.[34] Polls, obviously, are not clean representations of voting intentions. As such, to obtain a better estimate of the underlying process, I put them through a Kalman filter to obtain a single estimate per week (the average of all daily polls in that week). I choose a weekly frequency to strike a balance between limiting the need for interpolation and obtaining an acceptable number of observations for estimation. Weeks for which no poll is available are filled through linear interpolation and put through the filter using an arbitrarily large observation variance, which in practice means they are hardly taken into account for that week's estimated vote intentions. Given the filtered series represents the underlying vote intentions process, this represents a plausible approximation – vote intentions are likely to evolve relatively smoothly across time in between datapoints. Figure 6 depicts CDU voting intentions as evaluated through this method, with monthly frequency (Figure A2 presents the full timeframe version of this plot).

To add robustness to my analysis, I also provide an alternative interpolation method that takes the CDU communication strategy into account. Pereira (2019) distinguishes between two kinds of political party communication: valence-focused, emphasizing their own competence or integrity, or policy-focused, instead discussing their policy proposals. The author, using a dataset of newspaper articles conveying political agents' messages and public polls, finds that parties that over-perform on polls tend to focus more on valence issues, and that those that under-perform instead focus on policy issues. Given that parties have finer information on vote intentions than the general public,[35] this behavior can also be identified through Twitter and used to inform an interpolation process.

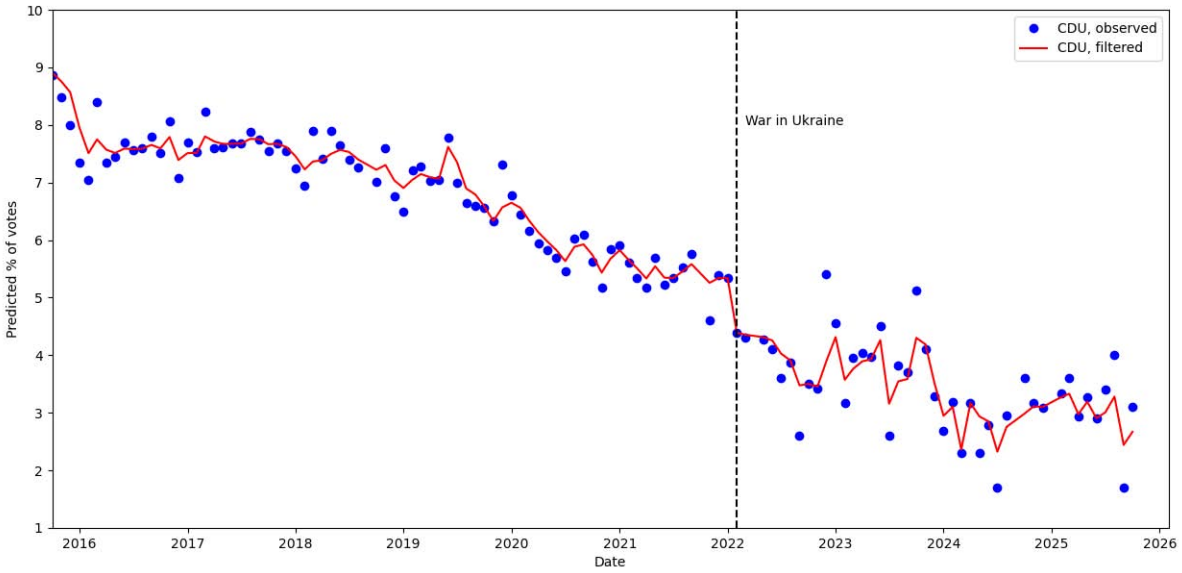
To this end, I use GPT-4.1 to classify each CDU tweet (comprising a total of 27 406 tweets) as policy-focused (P), valence-focused (V), or none (0). I run two separate zero-shot classifications using different prompts,[36] which are

[34]For days when more than one poll is released, I use an average value weighted with each poll's N. All polls can be found at https://www.marktest.com/wap/a/p/id~112/p~200909.aspx#t.

[35]Pereira (2019): "Nowadays, the vast majority of parties have private pollsters providing information that is not revealed to the public."

[36]The following parameters were used for both: max_tokens=1; temperature=0; system="You are a helpful research assistant.".

Figure 6: CDU predicted voting intentions



Notes: Observed and filtered monthly poll averages. All data is obtained from https://www.marktest.com/wap/a/p/id~112/p~200909.aspx#t.

compared in table 2. This table further presents performance metrics for this task, based on a manually-labelled set of 601 CDU tweets (meaning tweets by the CDU, PCP and PEV official accounts as well as by politicians affiliated with PCP or PEV).[37]

Several conclusions can be drawn from it: firstly, classification task #1 seems to perform better overall than #2. Secondly, both classifiers perform better for policy than for valence communication. Thirdly, they yield reasonably similar classifications, as indicated by Cohen's κ, more so when a binary task is considered (for which V and 0 are considered as the same classification). With all of these in mind, I opt to use the best-performing classifier – #1 – to inform a more complex interpolation based on the binary P/0 classification as follows:

$$\begin{cases} CDU_t = CDU_{t-1} + (CDU_N - CDU_\phi) \times \frac{1+pol_t}{\sum_{i=\phi+1}^{N}(1+pol_i)} & , CDU_N - CDU_\phi \leqslant 0 \\ CDU_t = CDU_{t-1} + (CDU_N - CDU_\phi) \times \frac{(1+pol_t)^{-1}}{\sum_{i=\phi+1}^{N}(1+pol_i)^{-1}} & , CDU_N - CDU_\phi > 0 \end{cases}$$

where t stands for each week, CDU_t is the missing poll datapoint to be interpolated, CDU_{t-1} is the previous poll observation (which might or not itself have been interpolated), CDU_N is the first actually-observed poll value after week t, CDU_ϕ is the last actually-observed poll value before week t (thus $\phi < t < N$), and pol_t corresponds to the number of policy-focused tweets made by CDU in week t. Briefly, it means that the change in CDU between

Prompt 1: "A given tweet by the CDU, PCP or PEV parties can be classified as valence- or policy-focused. Valence-focused tweets explicitly highlight the competence, strength, honesty, integrity, commitment or unity of the party or their representatives. Policy-focused tweets rather mention their own specific policy proposals or positions. Please label the following tweet (by CDU, PCP or PEV) as valence-focused (V), policy-focused (P), or neither (0). Avoid making assumptions, using neither (0) when the tweet refers exclusively to other parties' policies and when in doubt. Always begin your answer with the label you choose (V/P/0) and provide a short explanation. Tweet: {text}" Prompt 2: "Read the following tweet by the CDU, PCP, or PEV parties and classify it into one of three mutually-exclusive categories: V – Valence-focused: highlights competence, honesty, integrity, unity, leadership or other positive qualities of CDU/PCP/PEV or their members, without concrete policy content. P – Policy-focused: mentions specific policy proposals, programmes, measures, laws, or positions supported or opposed by CDU/PCP/PEV. 0 – Neither: does not focus on the party's own valence or policy or is unclear. If in doubt, choose 0. Output ONLY the single-character label V, P, or 0 — no explanation or extra text. Tweet: {text} Label:". Line breaks are omitted in this footnote for space savings.

[37] Figure A3, in the appendix, plots classification #1 across time.

Table 2: Valence-policy classifier metrics

		Precision	Recall	F1-Score	Support
#1	P	0.85	0.87	0.86	241
	V	0.64	0.62	0.63	85
	0	0.87	0.86	0.87	275
	V+0	0.91	0.90	0.90	360
	Accuracy (V/P/0)			0.83	601
	Accuracy (V/P+0)			0.89	601
#2	P	0.73	0.95	0.82	241
	V	0.80	0.47	0.59	85
	0	0.87	0.75	0.80	275
	V+0	0.96	0.76	0.85	360
	Accuracy (V/P/0)			0.78	601
	Accuracy (V/P+0)			0.84	601

	Multiclass (V/P/0)	Binary (P/V+0)
Cosine similarity	0.324	0.497
Cohen's κ	0.699	0.730

	=P	=V	=0
Full dataset (#1)	10 537	3 951	12 918
Full dataset (#2)	13 488	2255	11 663

weeks N and ϕ is distributed by each of the weeks between them according to the concentration of policy-focused tweets in each week: when CDU is decreasing, it decreases more in periods of higher policy-focused communication; while when CDU is increasing it increases less in periods of higher policy-focused communication.

In this alternative methodology, all weeks for which no real polls are available are interpolated according to the above rule and assigned an observation variance equivalent to an N=300 poll – slightly higher than what applies to the lowest-N polls in the dataset (for which N=400). This means these datapoints are not basically ignored, as happens in the first method; but still carry less weight on the filtered series than actual observations. Out-of-scope gaps, for which there is no Twitter data, are linearly interpolated and handled as before.

Figure 7 presents weekly CDU voting intentions using both linear and policy-weighted interpolation, after the application of the Kalman filter (figure A4, in the appendix, shows the entire 2010-2023 timeframe). It shows minimal differences between the two methods – nonetheless and for robustness, I will present results for both.

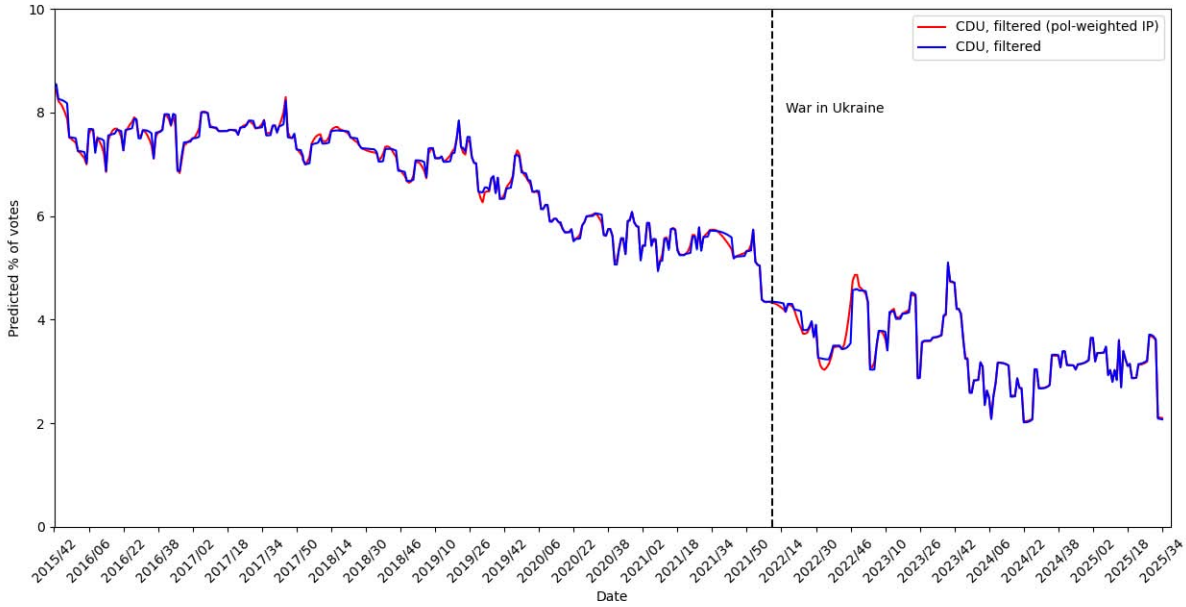
5 Inferential framework and results

> *@pcp_pt should be ashamed for supporting the war criminal, aggressor, and oppressor Putin and his minions like the dictator Lukashenko. And apologize to all Portuguese, Ukrainians, and Russians! Now! #SlavaUkraine!* [38]
>
> Ana Gomes (@AnaMartinsGomes), former presidential candidate (PS), February 24th 2022

With all data on hand and having the established the plausibility of the stigma discontinuity at the outbreak of the Russo-Ukrainian war, an inferential framework can be built. I propose two operationalizations of this natural experiment: a regression discontinuity design where I assume that the only relevant factor for CDU vote intentions

[38] Source: https://twitter.com/AnaMartinsGomes/status/1496743069034631170 (in Portuguese).

Figure 7: CDU predicted voting intentions (weekly)



Notes: Weekly poll averages, standard Kalman and pol-weighted interpolation.

that changed in February 2022 was the stigma stemming from the war; and a difference-in-differences approach where I assume that the only relevant factor for vote intentions that changed for CDU and not for BE was said stigma. Their particular design and implications are discussed in turn in what follows.

5.1 Regression discontinuity

I begin with a simple RDD framework, under the assumption that the only phenomenon which might have impacted vote intentions on PCP at the outbreak of the war was the widespread stigmatization they suffered. This is a plausible hypothesis, as the discussion in previous sections shows: PCP had for decades publicly held anti-NATO and -EU positions and specifically referred to the Ukrainian post-Maidan institutions as nazi-inspired.

I thus implement a framework where the PCP vote intention is the outcome variable and time is the running variable – by nature, there will be no risk of threshold manipulation. I do not include any controls and use both linear and quadratic fits in parametric and non-parametric estimations. Tables 3 and 4 present results, while figure 8 graphically depicts them.[39] Results – a negative, highly significant effect of stigmatization on vote intentions and no discernible effect on its growth – clearly suggest a level drop in the dependent variable without a significant change in trend compared to the pre-war period.

5.2 Difference in differences

While the RDD results are quite strong, one may argue that it is unlikely that the only factor that might have impacted PCP vote intentions in the third week of February 2022 was the increase in their stigmatization. Other potential relevant factors might have been their poor results in the January 30th elections or the renewed prominence of their anti-West stances. To curb these concerns, I propose a diff-in-diff approach using Bloco de Esquerda voting intentions as a control group.[40]

[39] Using the standard Kalman method. Using pol-weighted interpolation yields graphically-indistinguishable results.
[40] BE polls are treated using the standard Kalman method in all regressions.

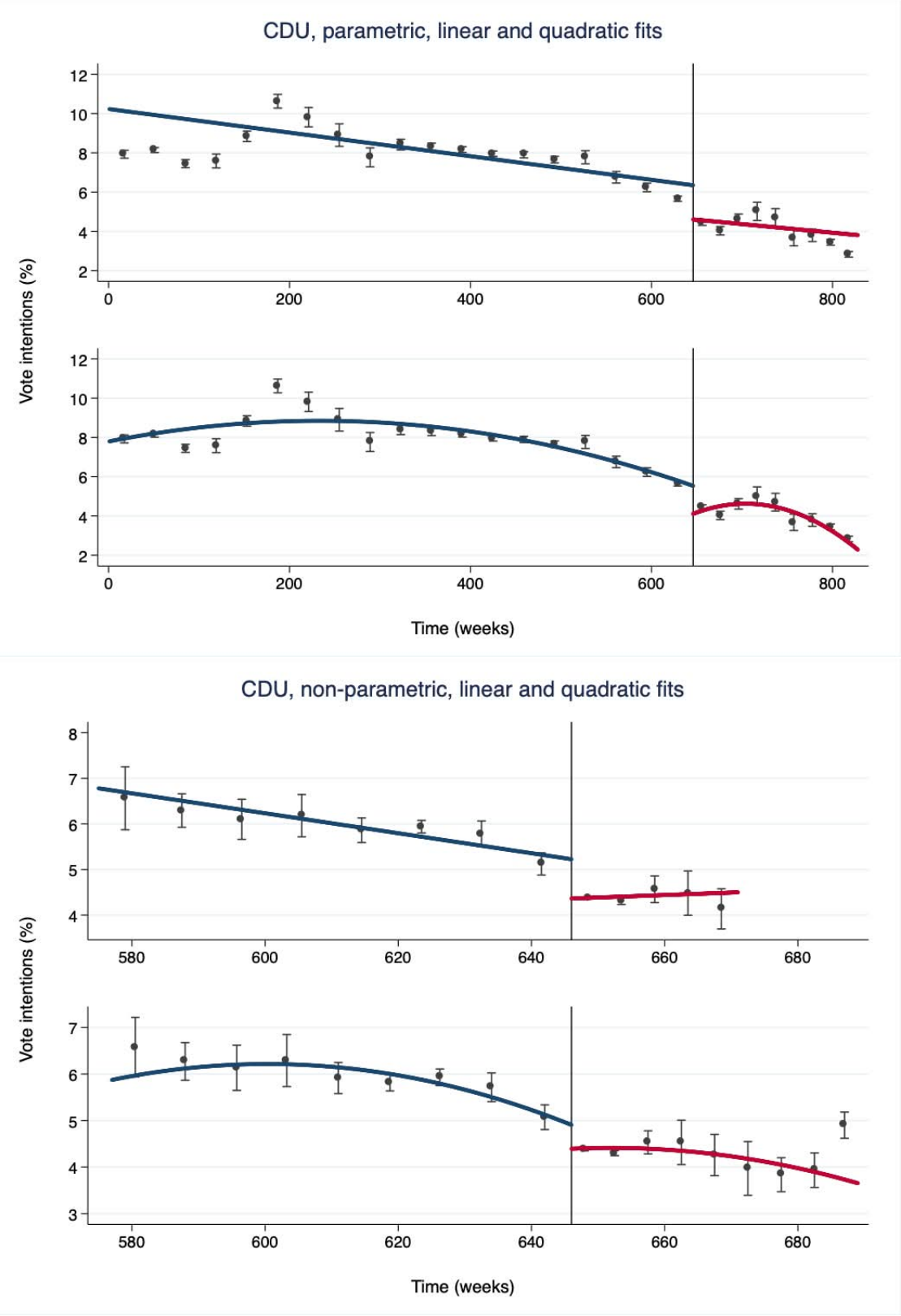
Table 3: RDD results, standard Kalman

$Y = Vote\,Intentions_{i,t}$					
	OLS	RD			
Treatment	-4.678	-0.566	-0.563	-0.485	-1.532
Robust p-value	0.000	0.001	0.002	0.000	0.000
Robust CI (95%)	[-4.853, -4.504]	[-0.898, -0.234]	[-0.918, -0.207]	[-0.670, -0.301]	[-1.777, -1.287]
Eff. # obs.	828	96	196	828	828
Control N		71	146	645	645
Treatment N		25	50	183	183
Polynomial order		1	2	1	2
Bandwidth selection		MSE-optimal		Parametric	
$Y = Vote\,Intentions\,growth_{i,t}$					
	OLS	RD			
Treatment	-0.009	-0.028	0.032	0.003	0.006
Robust p-value	0.740	0.962	0.598	0.953	0.909
Robust CI (95%)	[-0.063, 0.445]	[-0.118, 0.112]	[-0.087, 0.151]	[-0.087, 0.092]	[-0.104, 0.117]
Eff. # obs.	827	267	363	827	827
Control N		203	301	644	644
Treatment N		64	62	183	183
Polynomial order		1	2	1	2
Bandwidth selection		MSE-optimal		Parametric	

Table 4: RDD results, pol-weighted interpolation

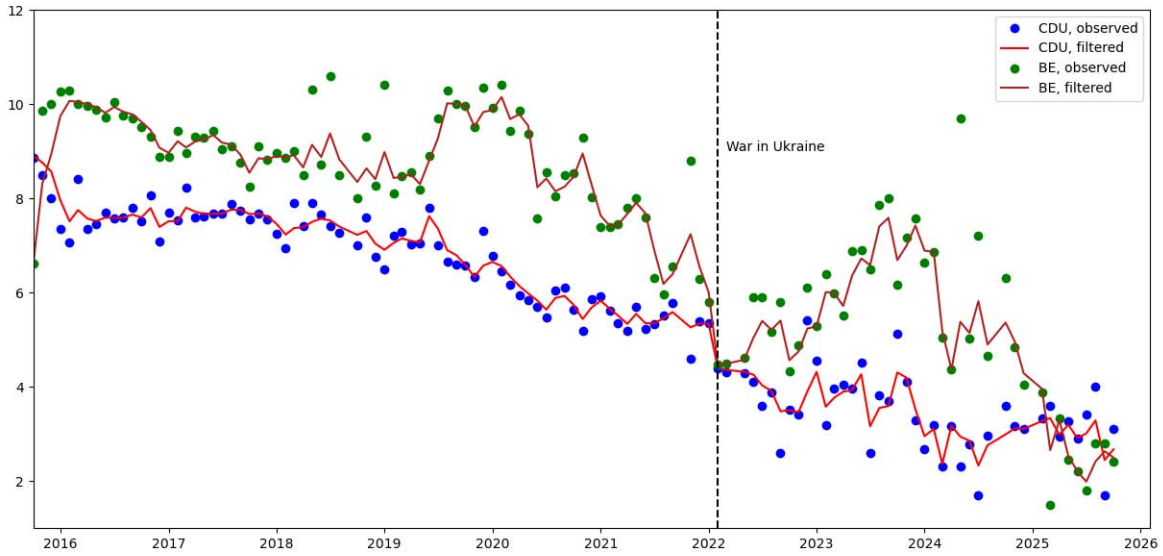
$Y = Vote\,Intentions_{i,t}$					
	OLS	RD			
Treatment	-4.678	-0.565	-0.570	-0.480	-1.525
Robust p-value	0.000	0.001	0.002	0.000	0.000
Robust CI (95%)	[-4.852, -4.504]	[-0.895, -0.236]	[-0.922, -0.217]	[-0.662, -0.298]	[-1.767, -1.284]
Eff. # obs.	828	96	196	828	828
Control N		71	146	645	645
Treatment N		25	50	183	183
Polynomial order		1	2	1	2
Bandwidth selection		MSE-optimal		Parametric	
$Y = Vote\,Intentions\,growth_{i,t}$					
	OLS	RD			
Treatment	-0.009	-0.003	0.031	0.003	0.006
Robust p-value	0.734	0.965	0.597	0.954	0.906
Robust CI (95%)	[-0.061 , 0.043]	[-0.114, 0.109]	[-0.085, 0.147]	[-0.084, 0.089]	[-0.100, 0.113]
Eff. # obs.	827	267	363	827	827
Control N		203	301	644	644
Treatment N		64	62	183	183
Polynomial order		1	2	1	2
Bandwidth selection		MSE-optimal		Parametric	

Figure 8: RDD results, parametric and non-parametric



Notes: Employed bandwidths are the same as in the estimations. Observations are bundled IMSE-optimal quantile-spaced bins.

Figure 9: CDU and BE predicted voting intentions, monthly



Notes: observed and filtered monthly poll averages.

Bloco de Esquerda are a good fit for this purpose for three reasons. Firstly, as figures 9 and 10 show, PCP and BE voting intentions followed similar trends since 2016.[41] Secondly, they were in the same position as PCP in the aftermath of the January 30th 2022 elections: both were the parliamentary support for the minority PS government; both did not approve the general budget, forcing a snap election; and both faced severe electoral losses compared to their previous results. Thirdly, they are also openly anti-NATO and anti-Western international institutions and their initial reaction to the Russian invasion was perceived as inconsistent – a course they quickly corrected, thus avoiding most stigma.

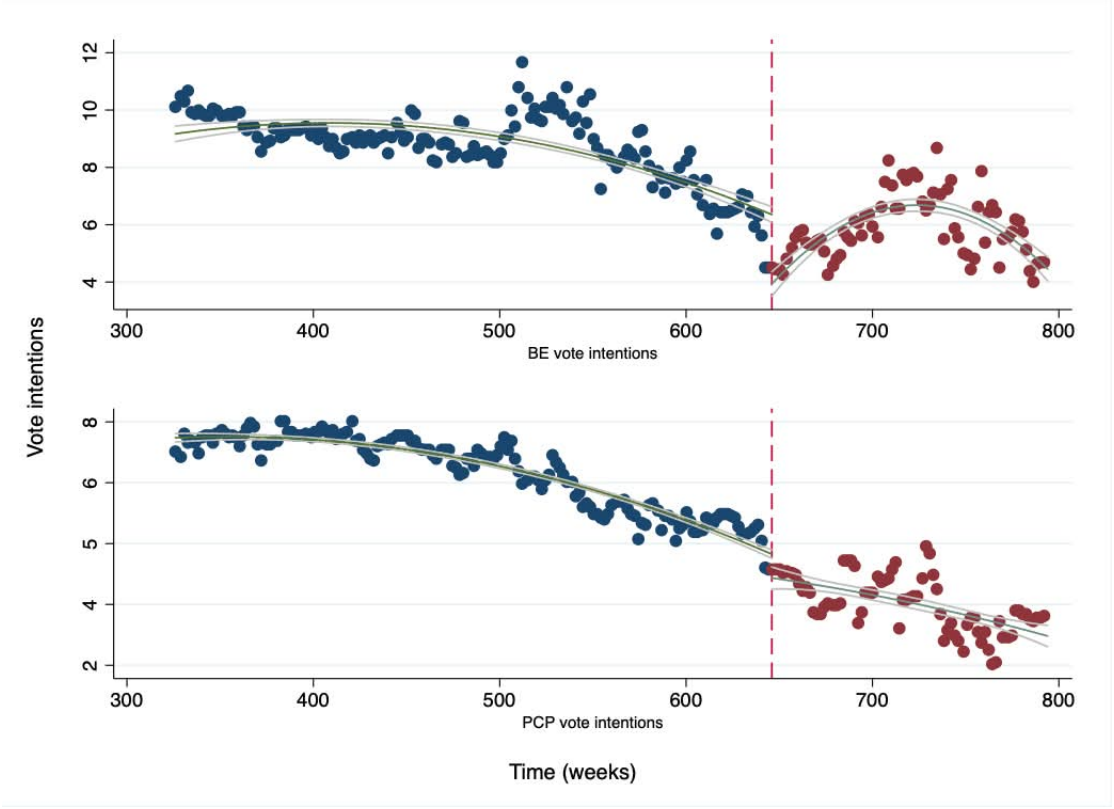
Specifically, my diff-in-diff framework is as follows.

$$VI_{i,t} = \alpha + \beta_1 Treat_i + \beta_2 War_t + \beta_3 (Treat_i \times War_t) + \varepsilon_{i,t}$$

Where $VI_{i,t}$ stands for the estimated vote intentions of party i (BE or CDU) at time t; $Treat_i$ is a dummy equal to zero if party i was not subjected to stigma (i.e. for BE) and 1 otherwise (i.e. for CDU), War_t is a dummy equal to 1 if week t is equal to or larger than 646 (standing for the third week of February 2022, when the Russo-Ukrainian war broke out) and zero otherwise; and $\varepsilon_{i,t}$ is an error term. For this exercise, I employ data from January 2016 until the end of 2024, which provides an ample window for the impact of the war to be taken into account and avoids BE's recent (and unrelated to this matter) electoral free fall. Table 5 presents results, suggesting that stigma had a negative and significant impact on both the level and growth of CDU vote intentions, until the end of 2024, when using BE as a counterfactual. Results are thus consistent across the RDD and DiD approaches regarding vote intentions; while only the DiD finds significant and consistent results for their growth. Whether one prefers the former or the latter depends on whether one believes pre-war CDU trends or post-war BE trends better represent how CDU vote intentions would have behaved post-war, which is up for discussion but difficult to objectively answer.

[41] Figure A5 presents the full 2010-2023 timeframe.

Figure 10: CDU and BE vote intentions with fitted curve, 2016-2024



Notes: observations are bundled into bins with 2 observations each.

Table 5: Diff in diff results

	$Y = Vote\,Intentions_{i,t}$		$Y = Vote\,Intentions\,growth_{i,t}$	
	(1)	(2)	(3)	(4)
War_t	-4.515	-4.529	0.216	0.220
	(1.204)	(1.185)	(0.106)	(0.111)
$Treat_i \times War_t$	-0.369***	-0.369***	-0.017***	-0.017***
	(0.001)	(0.001)	(0.001)	(0.001)
Interpolation	Linear	Pol-weighted	Linear	Pol-weighted
Fixed effects	Party, week	Party, week	Party, week	Party, week
R^2	0.907	0.907	0.485	0.486
Observations	938	938	938	938
Obs. per party	469	469	469	469

Notes: $Treat_i$ is omitted as it corresponds to the party fixed effect. Robust standard errors are employed, with values smaller than 0.001 rounded up to that value. *** stands for significance at the 1% level.

6 Concluding remarks

> *From Ukraine to Setúbal, to escape the Russian invasion. But in the Setúbal City Hall, led by PCP, they were received by Russians! Their documents are photocopied, and they ask the women where their husbands are! Their fear has returned! #SHAMEONPCP* [42]
>
> Pedro Frazão (@ODeputadoBanido), MP (CH), April 29th 2022

This work discusses the increasingly relevant phenomenon of political stigmatization, defined as the ostensive refusal by political agents to cooperate with or accept another party. This has become a relatively widespread tactic in representative democracies towards radical right parties, often revealing itself in cordon sanitaire and permanent, underlying political discrediting campaigns by a wide political majority towards a select few parties.

This raises several interesting questions, practical and otherwise. How does stigma impact the vote intentions of stigmatized and stigmatizing parties? What determines whether a given party is stigmatized or stigmatizes? In which cases is stigmatization a good practice for potentially-stigmatizing parties? When is this practice justified? Questions as these highlight the relevance of political stigmatization and its study, which will at least allow us to better understand what to expect from this practice and how it might fit into the democratic ideal.

The literature has answered this call in part – extensive research has been done on the electoral impacts of stigmatization on the radical right and on its determinants. There is, however, a limited amount of evidence regarding the stigmatization of left-wing parties, and the constant evolution of democratic polities calls for their continued study as different political forces interact, reshaping themselves and institutions.

As such, I have dedicated this work to discussing the particular case of the Portuguese Communist Party: until recently, an atypical case among modern communist parties of both electoral success and social acceptance, which has become the target of widespread stigmatization as a result of the Russian invasion of Ukraine (without undertaking any meaningful policy or communication change). This case offers an especially interesting case study with two main advantages: allowing a rare look at current-day stigmatization of a left-wing party and a natural experiment setting which makes the isolation of stigma effects especially plausible. Results indicate that the sudden wave of stigma had a significant, negative impact on PCP's vote intentions, with limited evidence of a similar impact on the growth of PCP vote intentions per period. This is generally in line with the findings of the literature regarding the right wing.

This paper provides several relevant contributions. Firstly, it adds a much-needed look at the radical left-wing to a literature almost exclusively focused on the radical right. Secondly, it implements a natural experiment framework that might be replicable in other polities. Thirdly, it offers a plausibly causal evaluation of an effect that is very hard to isolate, due to endogeneity and reverse causality issues. Finally, it delves into an especially interesting party in an especially relevant period of its existence, making for a potentially useful case study. Further work should aim to continue the expansion of the literature towards left-wing cases and look for other ways to isolate stigma effects.

References

[1] Akkerman, T., Rooduijn, M. (2015). "Pariahs or Partners? Inclusion and Exclusion of Radical Right Parties and the Effects on Their Policy Positions". Political Studies 63: 1140-1157.

[2] Art, D. (2007). "Reacting to the Radical Right: Lessons from Germany and Austria". Party Politics 13(3): 331-349.

[42] Source: unavailable (this account has been deleted and replaced by @Pedro_Frazao_). Original (in Portuguese): "Da Ucrânia até Setúbal, para fugir da invasão russa. Mas na Câmara de Setúbal, liderada pelo PCP, foram recebidos por russos! Os documentos são fotocopiados e perguntam às mulheres onde ficaram os maridos! O medo deles voltou! #VERGONHAPCP https://t.co/Phq2w5rv7W"

[3] Barberá, P., Casas, A., Nagler, N., Egan, P. J., Bonneau, R., Jost, J. T., Tucker, J. A. (2019). "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data". American Political Science Review 113(4): 883–901.

[4] Barberá, P., Rivero, G. (2015). "Understanding the Political Representativeness of Twitter Users". Social Science Computer Review 33(6): 712-729.

[5] Barberá, P., Steinert-Threlkeld, Z. C. (2019). "How to Use Social Media Data for Political Science Research". In L. Curini & R. Franzese (Eds.), The SAGE Handbook of Research Methods in Political Science and International Relations.

[6] Bischof, D., Wagner, M. (2019). "Do voters polarize when radical parties enter parliament?" American Journal of Political Science, 63(4): 888–904.

[7] Casas, A., Morar, D. (2015). "Different Channel, Same Strategy? Filling Empirical Gaps in congress Literature." In Proceedings of the Annual Meeting of the American Political Science Association, eds. Layna Mosley and Alvin Bernard Tillery. San Francisco: American Political Science Association. 1–21. Available at SSRN: https://ssrn.com/abstract=3395307.

[8] Cunha, A. (2020). "The Roots of the Portuguese Communist Party. The Introduction of Marxist Ideas in Portugal and the Creation of the Portuguese Maximalist Federation". Historia Contemporánea 64: 883-918.

[9] Ernst, N., Engesser, S., Büchel, F., Blassnig, S., Esser, F. (2017). "Extreme parties and populism: an analysis of Facebook and Twitter across six countries". Information, Communication & Society 20(9): 1347-1364.

[10] Fallaci, O. (1975, June 27). "Oriana Fallaci põe 'Cunhal a nu'". Jornal do Caso República, 6.

[11] Fazekas, Z., Popa, S. A., Schmitt, H., Barberá, P., Theocharis, Y. (2021). "Elite-public interaction on twitter: EU issue expansion in the campaign". European Journal of Political Research 60: 376–396.

[12] Han, K. J. (2020). "Reacting to Isolation: How the Political Exclusion of Extreme Right-wing Parties Changes the Party Support". Representation 56(1): 71-87.

[13] Kuran, T. (1987). "Preference falsification, policy continuity and collective conservatism". The Economic Journal, 97(387): 642–665.

[14] Lazar, M. (1988). "Communism in Western Europe in the 1980s". Journal of Communist Studies 4(3): 243-257.

[15] Lietz, H., Wagner, C., Bleier, A., Strohmaier, M. (2014). "When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter". Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media: 285-294.

[16] Lisi, M. (2008). "Rethinking the role of the Portuguese Communist Party in the transition to democracy". Portuguese Journal of Social Science 7(1): 17-35.

[17] Madeira, J. (2013). "História do PCP". Tinta da China.

[18] Meguid, B. (2005). "Competition Between Unequals: The Role of Mainstream Party Strategy in Niche Party Success". American Political Science Review 99(3): 347-359.

[19] Noelle-Neumann, E. (1993). "The spiral of silence: Public opinion - Our social skin (2nd ed.)." University of Chicago Press.

[20] Patrício, M. T., Stoleroff, A. D. (1994). "The Portuguese Communist Party: Perestroika and its Aftermath". In M. J. Bull & P. Heywood (Eds.), West European Communist parties after the revolutions of 1989. The Macmillan Press Ltd.

[21] Pereira, M. M. (2019). "Do parties respond strategically to opinion polls? Evidence from campaign statements". Electoral Studies 59: 78-86.

[22] Sassoon, D. (1992). "The Rise and Fall of West European Communism 1939-48". Contemporary European History 1(2): 139-169.

[23] Starobin, J. R. (1965). "Communism in Western Europe". Foreign Affairs 44(1): 62-77.

[24] Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. (2020). "The Dynamics of Political Incivility on Twitter". SAGE Open.

[25] Valentim, V. (2021). "Parliamentary Representation and the Normalization of Radical Right Support". Comparative Political Studies 54(14): 2475-2511.

[26] Valentim, V. (2022). "Political Stigmatization and Preference Falsification: Theory and Observational Evidence". Available at SSRN: https://ssrn.com/abstract=4023263.

[27] Valentim, V., Widmann, T. (2023). "Does Radical-Right Success Make the Political Debate More Negative? Evidence from Emotional Rhetoric in German State Parliaments". Political Behavior 45: 243-264.

[28] Van Spanje, J., Van Der Brug, W. (2007). "The Party as Pariah: The Exclusion of Anti-Immigration Parties and its Effect on their Ideological Positions". West European Politics, 30(5): 1022-1040.

[29] Van Spanje, J. (2010). "Parties beyond the pale: Why some political parties are ostracized by their competitors while others are not". Comparative European Politics 8(3): 354-383.

Appendix

A1. Parties explained

All parties mentioned in the text are briefly described below.

Partido Comunista Português (PCP): Founded in 1921, communists ("Portuguese Communist Party").

Partido Ecologista "Os Verdes" (PEV): Founded in 1982, ecologists. Not represented in parliament ("Ecological Party 'Greens'").[43]

Bloco de Esquerda (BE): Founded in 1999, radical left-wing ("Left Bloc").

Livre (L): Founded in 2014, eco-socialists ("Free").

Partido Socialista (PS): Founded in 1973, center left-wing ("Socialist Party").

Pessoas-Animais-Natureza (PAN): Founded in 2009, ecologists ("People-Animals-Nature").

Partido Social Democrata (PSD): Founded in 1974, center right-wing ("Social Democrat Party").

Iniciativa Liberal (IL): Founded in 2017, economic liberalism ("Liberal Initiative").

Centro Democrático Social - Partido Popular (CDS-PP): Founded in 1974, christian democrats ("Social and Democratic Center - People's Party").

Chega! (CH): Founded in 2019, radical and nationalist right-wing ("Enough!").

Juntos Pelo Povo (JPP): Founded in 2015, centrism and Madeira regionalism ("Together for the People").

Partido Popular Monárquico (PPM): Founded in 1974, monarchists. Not represented in parliament ("People's Monarchist Party").

Reagir Incluir Reciclar (RIR): Founded in 2019, centrist environmentalism. Never represented in parliament ("React Include Recycle").[44]

Movimento Democrático Português / Comissão Democrática Eleitoral (MDP/CDE): Founded in 1969 and dissolved in 1994, democratic socialism ("Portuguese Democratic Movement / Electoral and Democratic Commission").

Aliança Povo Unido (APU): Electoral coalition between PCP, PEV and MDP/CDE. Ran in the 1979, 1980, 1983 and 1985 legislative elections ("United People Alliance").

Coligação Democrática Unitária (CDU): Electoral coalition between PCP and PEV. First ran in 1987, ever-present in legislative elections since ("Democratic and Unitary Coalition").

Aliança Democrática (AD): Electoral coalition between PSD, CDS-PP and PPM. Ran in the 1979, 1980, 2024, and 2025[45] legislative elections ("Democratic Alliance").

Portugal à Frente (PAF): Electoral coalition between PSD and CDS-PP. Ran in the 2015 legislative elections ("Portugal Ahead").

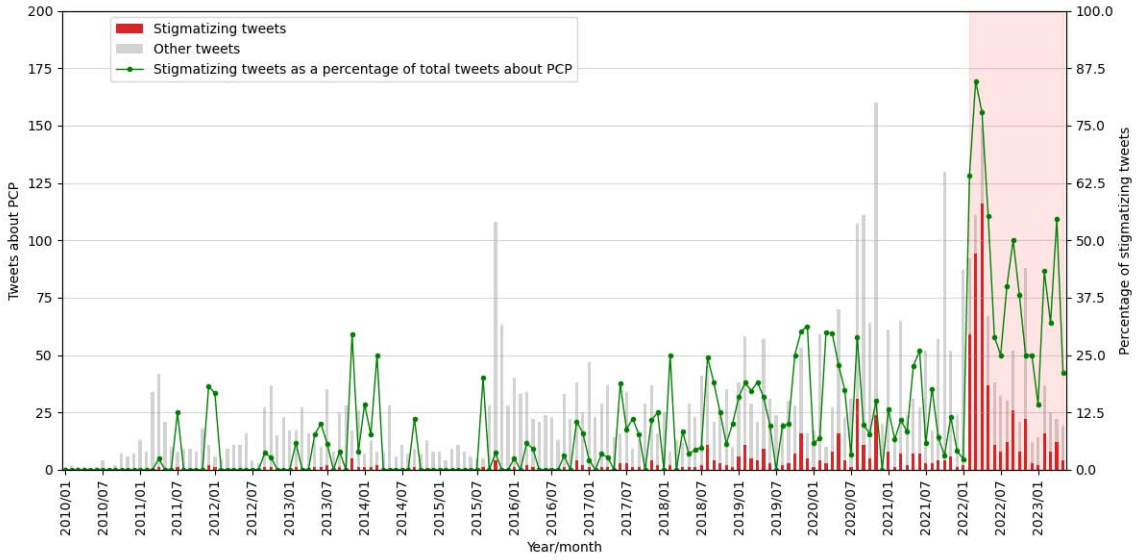
[43] Note that they have never run for office by themselves, but always as a part of the APU and CDU coalitions.

[44] The acronym reads "to laugh". This party was founded by Vitorino Silva, a TV personality and paver, to support his runs for parliament (2019, 2022 and 2024) and president (2021).

[45] PPM was not part of the coalition for the 2025 elections, but the AD naming was kept.

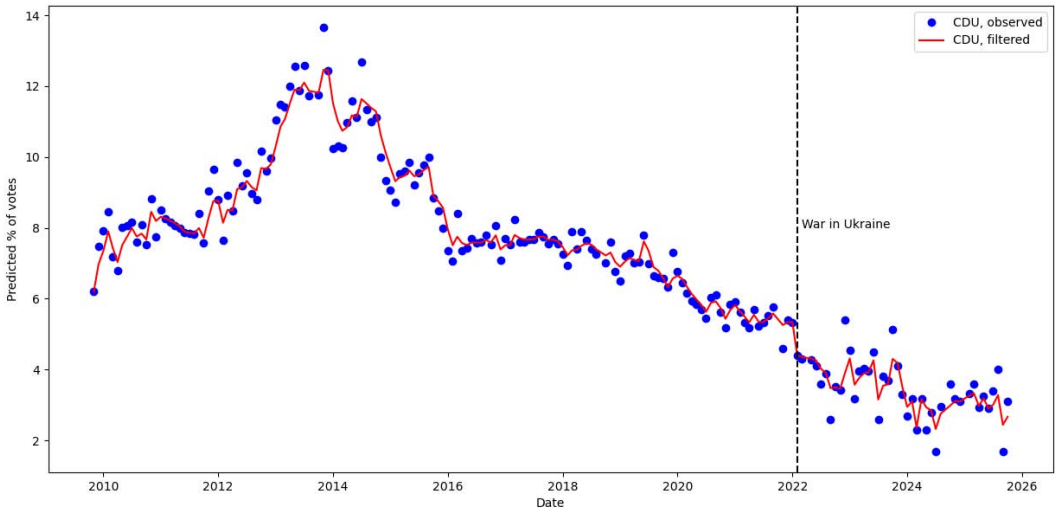
A2. Additional figures

Figure A1: Stigmatizing and non-stigmatizing tweets about PCP per month, 2010-2023



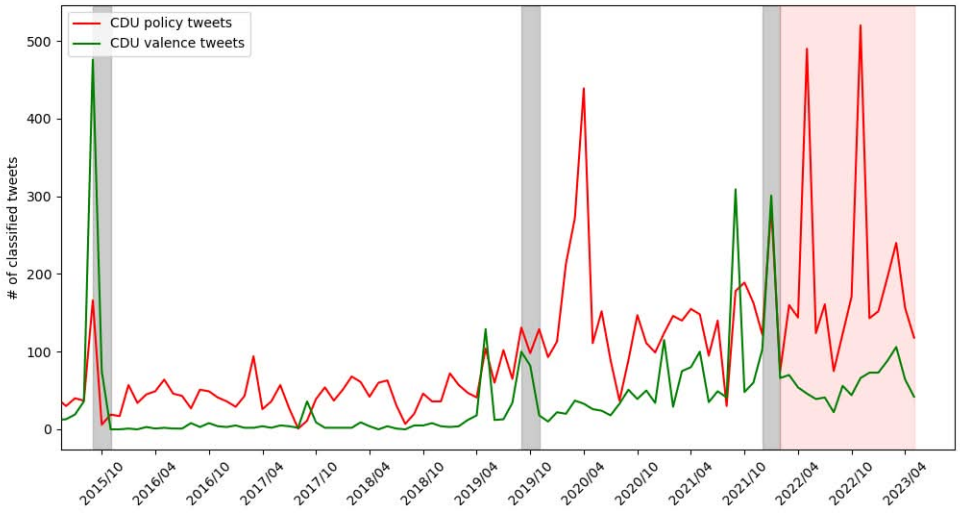
Notes: CDU, PCP and PEV tweets are excluded. The area shaded in red represents the war in Ukraine.

Figure A2: CDU predicted voting intentions, 2010-2023



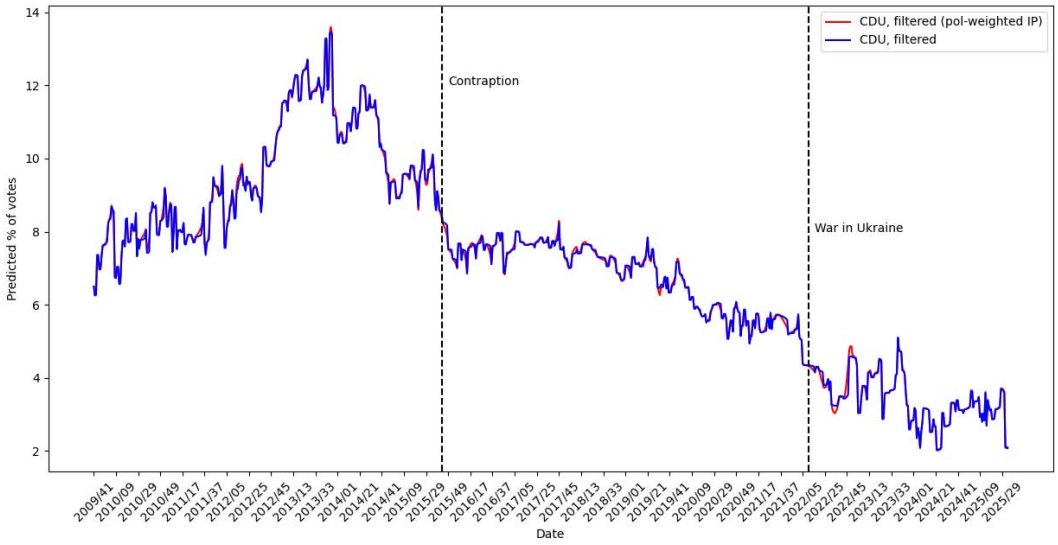
Notes: Observed and filtered monthly poll averages. All data is obtained from https://www.marktest.com/wap/a/p/id~112/p~200909.aspx#t.

Figure A3: Valence-policy classifier



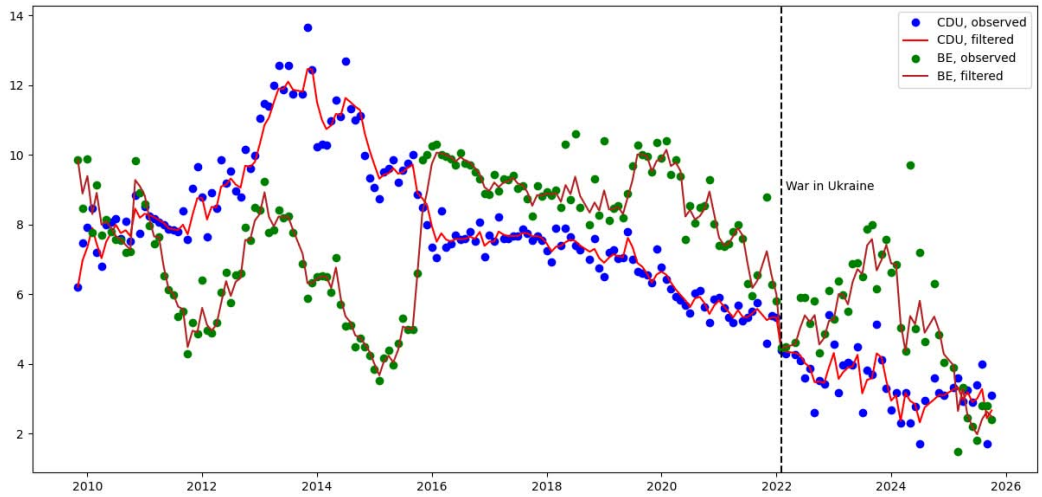
Notes: areas shaded in gray denote electoral campaign periods (3-month windows centered on the election month), while the area shaded in red represents the war in Ukraine.

Figure A4: CDU predicted voting intentions (weekly), 2010-2023



Notes: Weekly poll averages, standard Kalman and pol-weighted interpolation.

Figure A5: CDU and BE predicted voting intentions (monthly), 2010-2023



Notes: observed and filtered monthly poll averages.

A3. Included accounts/tweets

Notes: A PC role stands for presidential candidate, PM for the prime minister, and PAR for the president of parliament. Included are all MPs I could find on Twitter that served since 2015. Total tweets refer to the January 2010-May 2023 timeframe. Given that an account's handle can be changed, I provide the persistent Twitter account ID.

Name	Role	Party	Twitter ID	Total tweets	Observations
Mariana Rodrigues Mortágua	Leader	BE	1662057727	1884	Current leader
Beatriz Gebalina Pereira Gomes Dias	MP	BE	1371213404397040000	429	
Catarina Soares Martins	MP	BE	19472578	17471	Former leader
Isabel Cristina Rua Pires	MP	BE	3410400208	2688	
Joana Rodrigues Mortágua	MP	BE	572470115	4327	
João Manuel Duarte Vasconcelos	MP	BE	2957724905	1951	
Jorge Duarte Gonçalves da Costa	MP	BE	13695112	3119	
José Borges de Araújo de Moura Soeiro	MP	BE	1199624046067010000	208	
José Manuel Marques da Silva Pureza	MP	BE	20923305	465	
Luís Valentim Pereira Monteiro	MP	BE	879663520622804000	4345	
Maria Alexandra Nogueira Vieira	MP	BE	1103991204008650000	611	
Maria Manuel de Almeida Rola	MP	BE	97075632	165	
Moisés Salvador Coelho Ferreira	MP	BE	429013666	1416	
Nelson Ricardo Esteves Peralta	MP	BE	55435693	1373	
Pedro Filipe Gomes Soares	MP	BE	22007357	1976	
Sandra Mestre da Cunha	MP	BE	3009249448	1210	
Bloco de Esquerda	Party	BE	891843119238565000	3885	
Marisa Isabel dos Santos Matias	PC	BE	948552829	4309	
Francisco José Nina Martins Rodrigues dos Santos	Leader	CDS-PP	717487948128731000	282	Former leader
João Nuno Lacerda Teixeira de Melo	Leader	CDS-PP	2858866479	2150	Current leader
Maria de Assunção Oliveira Cristas Machado da Graça	Leader	CDS-PP	843122703594913000	465	Former leader
João Rodrigo Pinho de Almeida	MP	CDS-PP	1016799038182700000	538	
Telmo Augusto Gomes de Noronha Correia	MP	CDS-PP	60917604	91	
Centro Democrático Social - Partido Popular	Party	CDS-PP	19334929	6251	
Coligação Democrática Unitária	Party	CDU	3087068219	3298	
André Claro Amaral Ventura	Leader	CH	1097962618596320000	3033	Current leader
André Ventura (2)	Leader (2)	CH	1560621786781210000	14	Secondary account
Bruno Miguel de Oliveira Nunes	MP	CH	1195838239887750000	156	
Diogo Velez Mouta Pacheco de Amorim	MP	CH	22036916	0	Private account
Pedro Saraiva Gonçalves dos Santos Frazão	MP	CH	1508232014159300000	2463	
Rita Maria Cid Matias	MP	CH	1456351902703230000	397	
Rui Paulo Duque Sousa	MP	CH	48717561	566	
Chega	Party	CH	1070088307894340000	3006	
João Fernando Cotrim de Figueiredo	Leader	IL	1189235595341680000	491	Former leader
Rui Nuno de Oliveira Garcia da Rocha	Leader	IL	203263491	8005	Former leader
Bernardo Alves Martinho Amaral Blanco	MP	IL	712461001833254000	3115	
Carla Maria Proença de Castro Charters de Azevedo	MP	IL	1084384070686170000	4635	
Carlos Manuel Guimarães Oliveira Pinto	MP	IL	231703688	34407	
Joana Rita Madaleno Cordeiro	MP	IL	2786417932	1551	
Rodrigo Miguel Dias Saraiva	MP	IL	19722435	43898	
Iniciativa Liberal	Party	IL	2526916549	8332	
Tiago Pedro de Sousa Mayan Gonçalves	PC	IL	1286335166881960000	1292	
Rui Miguel Marcelino Tavares Pereira	Leader	L	14746456	21833	Current leader
Joacine Elysees Katar Tavares Moreira	MP	L	1214502815617420000	3961	
Livre	Party	L	2197487762	9271	
Paula Inês Alves de Sousa Real	Leader	PAN	616440530	2057	Current leader
Bebiana Maria Ribeiro da Cunha	MP	PAN	1248695290523150000	470	
Pessoas-Animais-Natureza	Party	PAN	51714305	7220	
Maria Cristina Pacheco Rodrigues	MP	PAN/CH	1337044292754350000	8343	
Alma Benedetti Croce Rivera	MP	PCP	1088454788394620000	981	
António Filipe Gaião Rodrigues	MP	PCP	21112266	897	
Bruno Ramos Dias	MP	PCP	759508476355874000	1644	
Duarte La Falher de Campos Alves	MP	PCP	1321243085897140000	1494	
João Guilherme Ramos Rosa de Oliveira	MP	PCP	21530275	157	
João Manuel Ildefonso Dias	MP	PCP	1155422878340460000	51	
Partido Comunista Português	Party	PCP	36392123	10120	
João Manuel Peixoto Ferreira	PC	PCP	951055588330475000	2021	
Os Verdes	Party	PEV	20978436	6743	
António Luís Santos da Costa	Leader	PS	1148538843458400000	531	Former leader
José Luis Pereira Carneiro	Leader	PS	1483224412006580000	412	Current leader
Pedro Nuno de Oliveira Santos	Leader	PS	727059918507892000	1397	Former leader
Ana Catarina Veiga dos Santos Mendonça Mendes	MP	PS	2743753215	699	
Ana Lúcia Silva de Passos	MP	PS	919980348896837000	984	
Ana Manuel Jerónimo Lopes Correia Mendes Godinho	MP	PS	752446483690090000	1213	
André Alexandre Pinotes Batista	MP	PS	22523863	8707	
Augusto Ernesto Santos Silva	MP	PS	1478796957082590000	652	
Carlos Filipe de Andrade Neto Brandão	MP	PS	20762598	5297	
Carlos João Pereira	MP	PS	4178367975	305	
Célia Maria Marques da Rosa Paz	MP	PS	1262056683540540000	1	
Cristina Maria da Fonseca Santos Bacelar Begonha	MP	PS	48706990	0	Never tweeted
Edite de Fátima Santos Marreiros Estrela	MP	PS	22904901	30600	
Eduardo Miguel Sabino Guedes Barroco de Melo	MP	PS	1119418973169370000	3868	
Elza Maria Henriques Deus Pais	MP	PS	825891155158134000	65	
Eurico Jorge Nogueira Leite Brilhante Dias	MP	PS	3125349375	510	
Eurico Jorge Nogueira Leite Brilhante Dias 2	MP	PS	1511087750698700000	76	
Fernando Medina Maciel Almeida Correia	MP	PS	2994469282	2704	

Name	Role	Party	Twitter ID	Total tweets	Observations
Francisco Miguel Vital Gomes do Vale César	MP	PS	21342208	1620	
Hugo Alexandre Polido Pires	MP	PS	4218594640	32	
Hugo Miguel Carvalheiro dos Santos Costa	MP	PS	21624322	504	
Hugo Miguel Costa Carvalho	MP	PS	1478798653536020000	28	
Isabel de Lima Mayer Alves Moreira	MP	PS	4278183399	1078	
Ivan Costa Gonçalves	MP	PS	1288111728585100000	345	
Joana Isabel Martins Rigueiro de Sá Pereira	MP	PS	1084788182200260000	33	
João Fernando Brum de Azevedo Castro	MP	PS	35721442	504	
João Nuno Ferreira Gonçalves Azevedo	MP	PS	31094851	34	
João Paulo de Loureiro Rebelo	MP	PS	751842311202541000	926	
João Paulo Moreira Correia	MP	PS	1189479354658700000	530	
João Saldanha de Azevedo Galamba	MP	PS	19713622	50566	
João Titternigton Gomes Cravinho	MP	PS	1690036087	2706	
João Veloso da Silva Torres	MP	PS	1536780608	643	
Jorge Filipe Teixeira Seguro Sanches	MP	PS	18717426	2255	
José Apolinário Nunes Portada	MP	PS	1155106755019460000	441	
José Carlos Ribeiro Barbosa	MP	PS	30921080	5044	
José Duarte Piteira Rica Silvestre Cordeiro	MP	PS	22874178	691	
José Manuel Santos de Magalhães	MP	PS	19265059	13603	
José Rui Alves Duarte da Cruz	MP	PS	1361070614614850000	17	
Lara Fernandes Martinho	MP	PS	4098156209	457	
Lúcia Fernanda Ferreira Araújo Silva	MP	PS	1319616495975040000	174	
Luís David Trindade Moreira Testa	MP	PS	30984457	723	
Luís Miguel de Freitas Marques Carvalho Soares	MP	PS	2996125256	30	
Mara Lúcia Lagriminha Coelho	MP	PS	321348534	2170	
Marcos da Cunha e Lorena Perestrello de Vasconcellos	MP	PS	1305536827810610000	106	
Maria do Céu de Oliveira Antunes	MP	PS	1113868486168730000	1606	
Maria Hortense Nunes Martins	MP	PS	48670023	11	
Mariana Guimarães Vieira da Silva	MP	PS	126118232	35669	
Mário José Gomes de Freitas Centeno	MP	PS	937770147925946000	35	
Marta Luísa de Freitas	MP	PS	1350413091587180000	0	Never tweeted
Miguel de Oliveira Pires da Costa Matos	MP	PS	942209900	2981	
Miguel dos Santos Rodrigues	MP	PS	789486947861467000	1448	
Miguel Filipe Pardal Cabrita	MP	PS	34908602	13387	
Nathalie Teixeira de Oliveira	MP	PS	539311525	552	
Nuno Jorge Cardona Fazenda de Almeida	MP	PS	1304392620974700000	0	Never tweeted
Olavo Balona Gouveia Câmara	MP	PS	1331592085745570000	11	
Pedro Carlos da Silva Bacelar Vasconcelos	MP	PS	1571494186548960000	16	
Pedro Filipe Mota Delgado Simões Alves	MP	PS	21642787	1025	
Pedro Miguel de Sousa Barrocas Martinho Cegonho	MP	PS	1313086116304740000	498	
Pedro Nuno Raposo Prazeres do Carmo	MP	PS	1246831029853310000	5	
Pompeu Miguel Noval da Rocha Martins	MP	PS	21784258	686	
Porfirio Simões de Carvalho e Silva	MP	PS	19564284	6989	
Raquel de Fátima Cardoso Ferreira	MP	PS	1388202635623290000	0	No text tweets
Ricardo Miguel Furtado Pinheiro	MP	PS	1349281332543810000	212	
Rita Mafalda Nobre Borges Madeira	MP	PS	1200887032312200000	240	
Sara Maria Belo Velez	MP	PS	51598436	26	
Sérgio Paulo Mendes de Sousa Pinto	MP	PS	290263087	534	
Susana Alexandra Lopes Correia	MP	PS	1279774531587010000	72	
Susana de Fátima Carvalho Amador	MP	PS	1351275867318700000	110	
Tiago Barbosa Ribeiro	MP	PS	21417835	5297	
Tiago Estevão Martins	MP	PS	2491663878	26	
Vera Lúcia Raimundo Braz dos Santos	MP	PS	1166094872341810000	6	
Augusto Ernesto Santos Silva	PAR	PS	1509952361355250000	652	
Partido Socialista	Party	PS	70349639	16716	
Ana Maria Rosa Martins Gomes	PC	PS	771383605	31283	
Primeiro Ministro	PM	PS	718009445863788000	5417	
Luís Filipe Montenegro Cardoso de Morais Esteves	Leader	PSD	1512019360147200000	279	Current leader
Rui Fernando da Silva Rio	Leader	PSD	1068848040834990000	1239	Former leader
Carlos Manuel Félix Moedas	Mayor	PSD	20058825	1872	President of the Lisbon Municipality
Alexandre Damasceno da Silva Poço	MP	PSD	187723641	2534	
Álvaro Fernando Santos Almeida	MP	PSD	31445468	447	
Ana Margarida Balseiro de Sousa Lopes	MP	PSD	40446161	4413	
Ana Miguel Marques Neves dos Santos	MP	PSD	3373604542	259	
André Guimarães Coelho Lima	MP	PSD	958255694708387000	1104	
António Alberto Maló de Abreu	MP	PSD	22263385	682	
António Manuel Pimenta Proa	MP	PSD	17477169	5196	
António Milton Topa Gomes	MP	PSD	1122554693752970000	98	
António Pedro Roque da Visitação Oliveira	MP	PSD	19252633	367	
Bruno Manuel Pereira Coimbra	MP	PSD	291761065	1	
Carlos Eduardo Vasconcelos Fernandes Ribeiro dos Reis	MP	PSD	1192552768671760000	708	
Catarina Leite de Faria da Rocha Ferreira	MP	PSD	21876067	927	
Cláudia Sofia Farinha André	MP	PSD	837087251020529000	586	
Cristóvão Duarte Nunes Guerreiro Norte	MP	PSD	608636651	930	
Duarte Filipe Baptista de Matos Marques	MP	PSD	21817509	30267	
Fernando de Carvalho Ruas	MP	PSD	959053700902899000	328	
Fernando Mimoso Negrão	MP	PSD	1395815076586720000	249	
Filipa Maria Salema Roseta Vaz Monteiro	MP	PSD	1021000641194450000	5	
Firmino José Rodrigues Marques	MP	PSD	1170677820114310000	1	
Hugo Daniel Alves Martins de Carvalho	MP	PSD	3395458955	208	
Hugo Miguel de Sousa Carneiro	MP	PSD	54363509	1706	
Hugo Patrício Martinho de Oliveira	MP	PSD	28378734	1740	
Isaura Maria Elias Crisóstomo Bernardino Morais	MP	PSD	1239250749001010000	3	
João Carlos Araújo Rêgo Montenegro	MP	PSD	1508374533211840000	88	
Joaquim José Miranda Sarmento	MP	PSD	880910811421712000	4844	
Jorge Paulo da Silva Oliveira	MP	PSD	1110871903374400000	74	
José Joaquim Cancela Moura	MP	PSD	1101549728880310000	156	
José Maria Lopes Silvano	MP	PSD	1071077491031320000	20	

Name	Role	Party	Twitter ID	Total tweets	Observations
Luís Manuel Morais Leite Ramos	MP	PSD	2450331027	282	
Márcia Isabel Duarte Passos Resende	MP	PSD	1195827393249850000	74	
Maria Emília e Sousa Cerqueira	MP	PSD	1230972281394000000	267	
Mónica Cláudia de Castro Quintela	MP	PSD	995433082588205000	536	
Nuno Miguel Oliveira de Carvalho	MP	PSD	1097787539379960000	86	
Paulo Cardoso Correia da Mota Pinto	MP	PSD	22820759	159	
Paulo César Rios de Oliveira	MP	PSD	1195000273288260000	4	
Pedro Nuno Mazeda Pereira Neto Rodrigues	MP	PSD	977250084554452000	7	
Ricardo Augusto Guerreiro Baptista Leite	MP	PSD	30839131	3844	
Ricardo Bastos Sousa	MP	PSD	1336258652953580000	7	
Sandra Cristina de Sequeiros Pereira	MP	PSD	1201547533	90	
Sara Martins Marques dos Santos Madruga da Costa	MP	PSD	852276914756751000	3	
Sofia Helena Correia Fernandes Sousa Matos	MP	PSD	1168985250439070000	49	
Tiago da Mota Veiga Moreira de Sá	MP	PSD	2461005242	1822	
Partido Social Democrata	Party	PSD	20560841	16640	
Vitorino Francisco da Rocha e Silva	PC	RIR	4644839074	722	
Disse Que Não Disse	Aggr		3193043925	-	Aggregator account for deleted tweets

CHAPTER TWO

Led by a green hand: Greta Thunberg, corporate tweeting and the stock market*

Henrique Alpalhão, Fabienne Röderer, José Tavares†

Abstract

Climate change activism has become increasingly relevant over the last decade, and has now an indisputable place of importance in social, political and economic discussion. These movements have given rise to several prominent figures – most notably Greta Thunberg – who, through their mass following and public interventions, arguably have the ability to influence markets, elections and other similarly grand outcomes. In this paper, leveraging an expansive Twitter (now X) dataset and LLM-powered text classification, we investigate how DAX40 corporate communication regarding climate change can affect their stock transactions and how Greta Thunberg might influence this link. Our results suggest that corporate outward communication on climate topics does influence the pricing of their stocks, and that alignment with Greta Thunberg is a relevant factor in this relationship: communicating contemporaneously with Thunberg is positively and significantly associated with stock price growth, but corporate and Thunberg's own climate communication are by themselves negatively associated with price growth. Based on these results, we propose that firms should make their public interventions on this topic in timely and parsimonious fashion: in the minefield of public climate communication, firms should be concerned less about thriving and more about surviving.

1 Introduction

> *Take that in. Then try wrapping your head around the fact that we're only seeing the beginning of a destabilising planet, while remembering that GHG emissions are still rising and the companies largely causing this are making record profits.*
>
> *A bit concerning, don't you think?*[1]
>
> Greta Thunberg (@GretaThunberg), July 23rd 2023

The topics of climate change, green activism and sustainable policymaking have, over the last decades, risen to the forefront of public discourse. Increasingly, social, economic and political actors are judged on their actions' environmental impact, with champions of the climate agenda rising to prominence. First and foremost of these

*This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab,Project 22209) and POR Norte (Social Sciences DataLab, Project 22209). This version: November 2025.

†Nova SBE and CEPR. E-mail: jtavares@novasbe.pt. Address: Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, 2775-405 Carcavelos, Portugal.

[1]Source: https://twitter.com/GretaThunberg/status/1683086734240645120.

individuals is Greta Thunberg who, in 2018, founded Fridays for Future (FFF) – a global student network against climate change which has, since then, promoted a sequence of global climate strikes. At their peak, in September 2019, and according to FFF's self-published data, these strikes spread to 3839 cities in 151 countries – a truly global phenomenon.[2]

Germany is one of the countries where climate change receives (and historically has received) more attention: in the aforementioned climate strike, 1.4 million protesters were reported in Germany alone.[3] On that day, a survey by the German public broadcaster found that 63% of respondents prioritized climate protection over economic growth. Younger, more educated and female respondents were further found to particularly favor fighting climate change.[4] When asked how to do so, 38% of Germans said it should be the responsibility of businesses and industry (European Commission, 2017).

Climate change generally refers to global warming on Earth. The term anthropogenic climate change is also used because, as most of the scientific community has found, it is partly human-induced through, amongst other things, greenhouse gas emissions (Houghton et al., 2001). It is through this angle of human culpability and responsibility to act that Greta Thunberg has contributed decisively to bringing climate change into the focus of public opinion, and has correspondingly achieved a sizable audience of 5.2 million followers on Twitter alone since joining in June 2018. This is especially true for Germany, which was second only to the USA in total number of school strikes until March 2023.[5] Discussion on climate change has been found to be highly emotional (Höijer, 2010); and the same can be said about Greta Thunberg herself, who has garnered both passionate support and opposition (Elgesem and Brüggemann, 2023; Park et al., 2021).

This strong and continuously-rising public attention towards climate issues is evident in international events such as the World Economic Forum or the UN Climate Change Conference (commonly known as COP). This phenomenon, however, has also spread online and is specifically quite relevant on Twitter, as Kirilenko and Stepchenkova (2014) and UN Global Pulse (2014) find. In the midst of this environment, corporations experience increased pressure and feel a responsibility to act (Levy, 2005; Porter and Kramer, 2011). Corporate climate action can be defined as companies integrating "environmental (...) concerns into their business operations and core strategy in close cooperation with their stakeholders" (European Commission, 2011, p.1). Broadly, it is a subcategory of Corporate Sustainable Responsibility (CSR) (Fernandez Gago and Nieto Antolín, 2004; Sdrolia & Zarotiadis, 2019), often taking the form of decarbonization measures (Sarasini and Jacob, 2014; Krabbe et al., 2015). This definition, covering all environmental concerns (e.g. reducing pollution, resource efficiency, the protection of nature), intersects well with Greta Thunberg's discourse.

In this work, we will delve into these topics by analyzing the potential link between corporate Twitter communication and stock market performance. Twitter provides an extremely granular tracker not only of Greta Thunberg's public interventions, but also of corporate ones – as Cho et al. (2016) find, companies increasingly turn to online platforms such as Twitter to communicate with their audiences. According to the Twitter Connect Playbook,[6] in 2022 Twitter boasted 229M monetizable daily active users.[7] Kirilenko and Stepchenkova (2014) state that Germany is the European country with the third most climate change-related tweets per day, and in Statista data Germany ranks 6th among the countries with most users with, as of February 2025, 21.63 million active accounts.[8]

[2]More information can be found at https://fridaysforfuture.org/. Note that more recent iterations of these strikes have gathered a remarkably lower number of people since the COVID-19 pandemic. These statistics can be found at https://fridaysforfuture.org/what-we-do/strike-statistics/list-of-countries/.

[3]Self-reported. Source: https://map.fridaysforfuture.org/list-towns.

[4]https://www.tagesschau.de/inland/deutschlandtrend-1807.html

[5]1798 in Germany vs. 3019 in the USA (self-reported). Source: https://fridaysforfuture.org/what-we-do/strike-statistics/list-of-countries/.

[6]https://business.twitter.com/en/resources/connect-playbook.html#connect-playbook-modal.

[7]"Twitter defines monetizable daily active usage or users (mDAU) as people, organizations, or other accounts who logged in or were otherwise authenticated and accessed Twitter on any given day through twitter.com, Twitter applications that are able to show ads, or paid Twitter products, including subscriptions." Source: https://investor.twitterinc.com/files/doc_financials/2022/q2/Final-Q2%2722-Earnings-Release.pdf.

[8]https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

This relevance notwithstanding, very little research has covered German Twitter data – as far as we are aware, no previous studies have attempted to capture corporate sentiment on climate change in Germany using it. We thus contribute to the literature in several ways: firstly by presenting an original approach to corporate communication in social media and the role of climate activists as opinion leaders; secondly by looking at the German case through an original, extensive Twitter dataset; and thirdly by attempting to connect Greta Thunberg's presence and action to corporate communication and stock market performance.

2 Literature review

Taking responsibility in a changing world[9]

Rheinmetall AG (@RheinmetallAG), Twitter biography

This section discusses relevant literature on climate change, corporate communication, their impact on stock markets, and the use of Twitter data. Greta Thunberg is also briefly introduced.

2.1 Climate change action of the DAX40 companies

Climate change has, in the 21st century, taken a central role in the agenda of political and business leaders. The 2015 Paris Climate Agreement,[10] however, was crucial in setting new carbon emission targets for participating countries (IPCC, 2013) and is the basis for several regulations that affect companies' strategies and financial returns (Tsalis and Nikolaou, 2017; Pham et al., 2019; Cadez et al., 2019; Lopes de Sousa Jabbour et al., 2020). To meet standards such as these, companies progressively adjust their strategies and operations (Weinhofer and Hoffmann, 2010; Downie and Stubbs, 2012).

Germany presents itself as a pioneer in climate protection and set successive sectoral targets towards greenhouse gas neutrality by 2050, as confirmed by the Ministry for the Environment's Climate Action Plan 2050.[11] Increasing regulation notwithstanding, market (e.g. employees, customers, and investors) as well as non-market stakeholders (e.g. activists, citizens, and governments) tend to demand climate-friendly practices beyond simple regulatory compliance (Delmas and Toffel, 2004; Reid and Toffel, 2009). Particularly in Germany, an increasing demand for sustainable finance led to the creation of a new Environmental, Social, and Corporate Governance (ESG) DAX50 Index in March 2020.[12]

Not all stakeholders, however, are equally powerful (Cadez et al., 2019). Under Freeman's stakeholder theory (Freeman, 1984), we expect that a firm's stance towards climate change is motivated by the pressure and demands of its stakeholders (Fieseler et al., 2010; Cadez et al., 2019; Zhang and Zhu, 2019). Depending on organizational characteristics and stakeholder interests, thus, companies participate in governmental programs, comply with voluntary standards of non-governmental organizations (NGOs) or directly interact with customers and suppliers (Delmas and Toffel, 2012). Moreover, companies respond to stakeholder pressure with green innovations, both product- and process-related (Zhang and Zhu, 2019).

Firms that adjust their overall strategy have been found to be successful in reducing emissions (Banerjee, 2002; Cadez et al., 2019). To reap the reputational rewards that most motivate corporate climate action, however, stakeholders must be informed – the literature on sustainability reporting looks into this. Firstly, according to Castelo Branco and Lima Rodrigues (2008), annual reports tend to present a higher number of environmental topics than corporate webpages (the opposite happens for other CSR topics). On the specific case of Germany, Kilian and

[9] Source: https://twitter.com/RheinmetallAG.
[10] For more information on the agreement, see https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.
[11] The English version of which can be found at https://www.bmuv.de/fileadmin/Daten_BMU/Pools/Broschueren/klimaschutzplan_2050_en_bf.pdf
[12] https://stoxx.com/index/daxesgk/.

Hennigs (2011) review the annual reports of DAX30[13] companies for CSR activities between 1998 and 2009. They find that, for environmental topics, stakeholder- and value-oriented communication had increased in contrast to performance-oriented communication, suggesting effective stakeholder pressure on firms. These results also reflect the increasing relevance of sustainability reporting, although no uniformity among companies was observed (Greiwe and Schönbohm, 2011; Dietsche et al., 2019).[14]

2.2 Twitter and its importance in social sciences

Twitter (now X) is a social media platform used by private and institutional users to exchange views on politics, news, and trends with a global community in the form of short messages (called tweets). The evolution of its writing prompt, from the original "What are you doing?" to the current "What's happening?", illustrates Twitter's crucial role in the evolution from print media to internet-based news portals. While users were previously encouraged to focus on private updates, the new question prompts them to report on their perception of their environment (Ebner et al. 2010). Additionally, the maximum number of characters per tweet has increased over time: from an initial 140 characters to 280 in 2017; and recently, for Premium users in the X era, to 25 000 characters in June 2023.[15] This provides at least twice as much space for content as initially for all users, and a much longer format for those paying.

This evolution offers additional opportunities for companies, which use the internet to connect with all kinds of stakeholders. On Twitter, companies can passively seek information regarding what stakeholders care about; but also actively shape their reputation through corporate communication, namely with regard to their environmental actions (Fieseler et al., 2010). 96% of the Fortune 500 companies manage an active Twitter account (Barnes et al., 2020). Having identified a positive effect of Twitter engagement on overall sales, Barnes and Lescault (2013) recommend that companies focus their social media activities towards it. Unlike companies themselves, CEOs tend to be less present on Twitter; the ones that are, however, achieve high stakeholder interaction (Capriotti and Ruesja, 2018). This provides some evidence of stakeholder demand for engagement on Twitter.

Additionally, what happens on Twitter has been found to have an impact on the stock market. Firstly, Gomez-Carrasco and Michelon (2017) find that tweets from consumer associations and trade unions have a significant and negative influence on stock prices and trading volume. Furthermore, several studies investigate the public discussion of climate change on Twitter. In 2014, the UN Global Pulse estimated that 140,000 English-language climate change-related tweets were made daily, with the peak in positive sentiment occurring on June 5th, the World Environment Day (UN Global Pulse, 2015). Dahal et al. (2019), on the other hand, carried out a volume and sentiment analysis as well as topic modeling to evaluate public climate change tweets. Their results suggest a negative global sentiment, especially if tweets co-occur with political events such as the United States of America withdrawal from the Paris Agreement or extreme weather events. They further note a lack of consensus on climate change issues in different geographical areas: in Australia the most discussed topic is energy; in the United Kingdom it is the ecological footprint; and in the USA hurricanes are most talked about. Koenecke & Feliu-Fabà (2019), further, find that extreme weather events increase the amount of climate change affirming tweet sentiment in the USA; Pearce et al. (2014) conclude that, in Australia, climate change affirming users communicate more with likeminded users. Leas et al. (2016), finally, obtain results that suggest that a single person can influence the conversation about climate change on Twitter: the DiCaprio effect describes an abnormal and significant increase in climate-related tweets after actor Leonardo DiCaprio had mentioned climate change in his speech at the 2016 Oscars.

[13] The previous version of DAX40, which included only 30 corporations.

[14] Note that previous results on DAX40 companies must be viewed with caution as index membership varies over time – its current composition will be presented in the coming sections.

[15] https://twitter.com/Write/status/1674221120495685635

2.3 Greta Thunberg and Thunberg effects

Greta Thunberg is a Swedish climate activist. In August 2018, at the age of 15 and amidst the then-hottest recorded summer in Sweden, she began a trend of "climate strikes" by skipping school on Fridays (Thunberg, 2019). In the wake of this, Thunberg garnered broad media coverage and inspired students across the world to join her campaign, under the hashtag "#FridaysForFuture". Today, Fridays for Future represents an international youth movement, with a self-reported, accumulated about 194 000 school strikes and 18 million strikers in 234 countries under their belt.[16]

Thunberg's audience, however, is not limited to the younger generations. On Twitter, where she garners intense and polarized reactions (Jung et al., 2020), influential accounts mention her to underscore and emphasize their own positions and opinions. Examples include media channels like The New York Times Media or BBC News Media, political leaders such as Barack Obama, private sector executives such as Elon Musk and civil rights activists (Jung et al., 2020).

The internet can function as a powerful platform for previously-unheard and financially underprivileged groups to reach broad audiences (Amichai-Hamburger et al., 2008). Greta Thunberg has taken advantage of this to great effect, having accumulated over 5 million followers since joining Twitter in 2018.[17] She was further elected Time Person of the Year 2019 for creating a global climate change global movement[18] and nominated for the Nobel Peace Prize multiple times. Her speeches at conferences such as the UN Climate Summit achieved global reach, with excerpts such as "How dare you?" or "Our house is on fire" obtaining significant notoriety (Jung et al., 2020).

This reach has led to several instances of so-called "Greta effects". In 2019, the Economist reported that Swedes, inspired by Thunberg, had minimized the number of domestic flights they took.[19] That same year, the Guardian found that British children were engaging more in activism through social media and that businesses were investing four times more in carbon offsetting to reduce their climate impact.[20] Sabherwal et al. (2020), finally, find that being familiar with Greta Thunberg positively predicts interest in climate action, labelling it the "Greta Thunberg Effect".

On January, 2020, Thunberg went a step further by actively asking Siemens to shut down a project due to its negative climate impact.[21] The company responded by offering a German representative of Fridays for Future a position in their supervisory board (which was rejected). Later that month, she attended the World Economic Forum (WEF) in Davos for the second time in a row, where she called on companies to take climate change seriously and stop using fossil fuels. These events show that not only is Thunberg specifically interested in corporate activity and its climate impact, but also that corporations take the bad publicity she might bring seriously.

3 Data

> *Driving decarbonization and digitalization. Together.*[22]
>
> Infineon Technologies AG (@Infineon), Twitter biography

This section describes our research question, proposed mechanism, used data and its manipulation, and the empirical strategy we employ. The purpose of this work is to investigate the potential effect of Greta Thunberg's

[16] These and more statistics can be found at https://map.fridaysforfuture.org/lists.
[17] https://twitter.com/GretaThunberg. Compared to her, for example, 2014 Nobel Peace Prize laureate and female rights advocate Malala Yousafzai (who also started her activism as a teenager) has so far only accumulated 1.8 million Twitter followers (https://twitter.com/Malala).
[18] https://time.com/person-of-the-year-2019-greta-thunberg/.
[19] https://www.economist.com/graphic-detail/2019/08/19/the-greta-effect.
[20] https://www.theguardian.com/environment/2019/nov/08/greta-thunberg-effect-driving-growth-in-carbon-offsetting.
[21] https://twitter.com/gretathunberg/status/1215919031494070272.
[22] Source: https://twitter.com/Infineon.

tweeting behavior on corporate outcomes, namely their own tweeting behavior and stock market performance. We are also interested in potential links between different responses to Thunberg and different market outcomes – from a market perspective, should firms engage in public climate debate or remain silent? This is based on the idea that investors include corporate social media presence and an environmentalist agenda (or at least environmental communication) in the group of relevant factors for investment decisions, which the literature has backed – see, for example, Müller et al. (2023), whose results suggest a strong link between social media content and interactions and personal investment decisions.

To lay the groundwork for this analysis, we begin by exploring the variations in corporate tweet volumes across our timeframe and their correlation with Thunberg's activity. We then employ regression analysis to evaluate how corporate tweeting on environmental topics might impact each company's stock.

This speaks to a limitation of the literature that has been frequently identified: the unclear impact of communicating climate change topics to stakeholders (Castelo Branco and Rodrigues, 2008; Kilian and Hennigs, 2011; Cadez et al., 2019). While annual reports and corporate webpages might capture mostly investors and the financial community, social media enables organizations to interact with a wider range of stakeholders (Kaplan and Haenlein, 2010). By using Twitter data, we attempt to capture a broad, mixed consumer-investor audience that is likely to influence corporate decision-making.

3.1 Data collection, processing, and initial analysis

To undertake this analysis, we employ two kinds of daily data: tweet counts and stock market performance indicators. We collect these series for German firms listed in the DAX40 index, as of 2023: the 40 largest German companies, which together account for around 80% of Germany's stock market capitalization.[23] In line with the literature (e.g. Bowen, 2020), we expect these companies to be especially visible (i.e. more reported on by the media), and thus potentially more active on Twitter than others. With greater media observation, stakeholder pressure intensifies and therefore large companies tend to respond more actively to climate change issues (Sharma and Nguan, 1999; Castelo Branco and Rodrigues, 2008; Morales-Raya et al., 2019). Due to the leading role of these companies, however, their analysis is also expected to provide information on far-reaching trends in the corporate use of social media for climate change communication.

We begin by collecting all tweets made by Greta Thunberg and all DAX40 companies that have a Twitter account, in the 2013-2023 period (table 1, further below, describes this set). Given Thunberg joined in June 2018, this timeframe captures 5 years of corporate activity where she was not a relevant actor and another 5 where she held significant sway over public opinion. For this task, the 'academictwitterr' R package was used (Barrie and Ho, 2021). Note that, at the time of writing, academic access to the Twitter API has been discontinued, compromising the use of this package and the mass collection of tweets in general. Our dataset was built previous to this and runs until May 2023.

We obtain the DAX40 composition from Börse Frankfurt.[24] Twitter accounts for each firm were collected through an organic Twitter search and official company websites. Not all companies have an account – specifically and as of May 2023, Hannover Re and Porsche SE do not seem to be present on the platform, and are as such excluded from our sample.[25] Although all other firms have Twitter accounts, not all are officially verified – where necessary and possible, we have verified their legitimacy through official corporate webpages. When several accounts exist, we select the ones directed at the world market and/or those that tweet in English.[26] To facilitate and homogenize

[23] For more information on the index, see https://stoxx.com/index/DAX/.

[24] https://www.boerse-frankfurt.de/indices/dax/constituents.

[25] Three possible accounts exist for the first – @hannover_re, @HNR_Gruppe and @hannover_rueck. The first two have tweeted last in December 18th 2012, and as such are not present in our timeframe, the third has never tweeted. Several Porsche accounts exist, but none seems to correspond to Porsche SE.

[26] For Example: @DeutscheBank tweets in English, while @DeutscheBankAG tweets mostly in German.

the analysis of tweets in languages other than English, we employ GPT 4o-mini to translate them into English prior to our main analysis, so as to avoid any selection issues in our sample. This was applied to 115 125 non-English language tweets in total, the distribution of which can be found in table 1.

All tweet text was then cleaned and processed using both a supervised machine learning classifier and a zero-shot GPT-based classifier that identify whether each tweet refers to climate change or not. The machine learning classifier is based on a set of manual labels, collected in three waves, from which the classifier learns to then apply the same method to the entire dataset. For the first wave, 2000 tweets were randomly selected from those that mentioned at least one of the following terms: "carbon"; "climate"; "environment"; "fossil"; "global warming"; "green"; "oceans"; "cop24"; "cop25"; "cop26"; "cop27"; "davos"; "ipcc"; "wef"; "world economic forum"; "sustainable"; "sustainability"; "greta thunberg"; "greta"; "thunberg"; "fridays for future"; "fridaysforfuture"; "circular"; "earth"; "earthday"; "plastic"; "recycle"; and "renewable". The second wave consisted of 1500 fully random tweets. The final wave consisted of 1000 tweets by Siemens and Covestro specifically, motivated by the fact that, when using a dataset of only the first two waves, our classifier returned an atypically high number of environment-related tweets for these two companies. When using the full manual classification dataset, this phenomenon is still present, but attenuated to a degree that is more plausible. By the end of this process, and after dropping duplicates, we obtain a learning set of 3 603 tweets, of which 1 741 do not refer to the environment and 1 862 do.[27]

The GPT classifier, on the other hand, employs GPT 4o mini and is a zero-shot task, prompted with the following message, followed by each tweet:

> "Consider the following tweet and classify it as being about climate change or not. Tweets about climate change might mention it directly or discuss sustainability concerns, fossil and renewable energy, activists like Greta Thunberg or events like the COP or WEF. Respond only with '1' for yes or '0' for no. Text:"

Table A1, in the appendix, reports accuracy metrics for the two classifiers. While both turn out to be quite precise, the GPT classifier surpasses the ML one by achieving a 93% overall F1 score. Given this, we choose to employ the former.

The final step in this process is to apply the classifier to the entire dataset. When done, this yields a split of 29 671 corporate tweets about the environment (out of a total of 509 149) and 9 556 Greta Thunberg tweets about the environment (out of a total of 10 998). Table 1 presents all firms and individuals included in our analysis and initial descriptive statistics, while figure 1 presents tweet distribution by firm as a stacked bar chart. From it, we can immediately conclude that our database is significantly unbalanced in terms of firm participation: the top 4 most active firms, Mercedes-Benz, Deutsche Telekom, BMW, and Adidas, account for almost 220 000 tweets – approximately 40% of the sample. The intensity of environment-related tweeting, however, does not seem to be predicted by overall tweeting amount.

4 Methodology

> *We supply material solutions for the great challenges of our time. Follow us on our journey to becoming #FullyCircular.*[28]
>
> Covestro AG (@covestro), Twitter biography

In looking for an effect of climate tweeting on the stock market, it is wise to begin by identifying the specific mechanisms through which it might occur. We have already seen the literature establish that stakeholders want

[27] These figures include 429 Greta Thunberg tweets, which we have included in the learning set as they are very dense in environment-related terms. If they are removed from the set, we obtain 1 439 tweets about the environment and 1 735 about other topics.

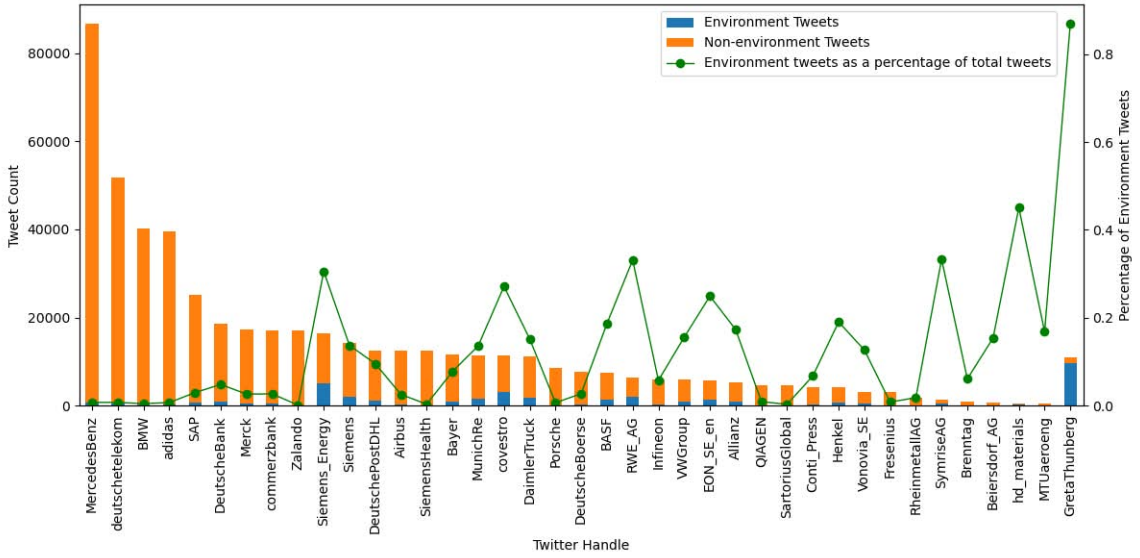
[28] Source: https://twitter.com/covestro.

Table 1: Included entities and descriptive statistics

Entity	Ticker	Twitter	# env. tweets	# tweets	% translated
Adidas AG	ADS.DE	adidas*	276	39470	2.31%
Airbus SE	AIR.DE	Airbus*	315	12450	5.45%
Allianz SE	ALV.DE	Allianz*	908	5241	4.22%
BASF SE	BAS.DE	BASF	1377	7373	3.88%
Bayer AG	BAYN.DE	Bayer*	907	11677	15.81%
Beiersdorf AG	BEI.DE	Beiersdorf_AG	101	656	7.16%
BMW AG St	BMW.DE	BMW*	188	40141	2.62%
Brenntag SE	BNR.DE	Brenntag	59	968	5.06%
Commerzbank AG	CBK.DE	commerzbank*	455	17034	87.72%
Continental AG	CON.DE	Conti_Press; Conti_Industry	286	4221	14.93%
Covestro AG	1COV.DE	covestro*	3067	11303	1.42%
Daimler Truck Holding AG	DTG.DE	DaimlerTruck	1707	11233	19.65%
Deutsche Bank AG	DBK.DE	DeutscheBank	892	18488	8.21%
Deutsche Börse AG	DB1.DE	DeutscheBoerse	209	7668	19.80%
Deutsche Post AG	DHL.DE	DeutschePostDHL*	1200	12542	38.54%
Deutsche Telekom AG	DTE.DE	deutschetelekom*	385	51791	96.45%
E.ON SE	EOAN.DE	EON_SE_en*	1441	5782	2.20%
Fresenius SE & Co. KGaA	FRE.DE	Fresenius	26	3176	39.45%
Hannover Rück SE	HNR1.DE				
Heidelberg Materials AG	HEI.DE	hd_materials	209	464	8.84%
Henkel AG & Co. KGaA Vz	HEN3.DE	Henkel	786	4126	12.75%
Infineon Technologies AG	IFX.DE	Infineon	344	5980	19.10%
Mercedes Benz Group AG	MBG.DE	MercedesBenz	645	86710	2.76%
Merck KGaA	MRK.DE	Merck*; merckgroup*	456	17338	1.61%
MTU Aero Engines AG	MTX.DE	MTUaeroeng	69	410	42.68%
Münchener Rück AG	MUV2.DE	MunichRe	1542	11338	6.74%
Porsche AG Vz	P911.DE	Porsche*	56	8491	5.46%
Porsche Automobil Holding SE	PAH3.DE				
Qiagen N. V.	QIA.DE	QIAGEN	43	4678	0.88%
Rheinmetall AG	RHM.DE	RheinmetallAG*	34	1877	25.63%
RWE AG St	RWE.DE	RWE_AG	2108	6361	89.06%
SAP SE	SAP.DE	SAP	747	25047	2.52%
Sartorius Ag Vz	SRT3.DE	SartoriusGlobal	14	4646	1.98%
Siemens AG	SIE.DE	Siemens*	1939	14147	6.88%
Siemens Energy AG	ENR.DE	Siemens_Energy*	5014	16405	3.71%
Siemens Healthineers AG	SHL.DE	SiemensHealth*	43	12400	0.90%
Symrise AG	SY1.DE	SymriseAG	488	1465	3.89%
Volkswagen Group AG Vz	VOW3.DE	VWGroup	909	5844	13.26%
Vonovia SE	VNA.DE	Vonovia_SE*	404	3191	98.06%
Zalando SE	ZAL.DE	Zalando	22	17017	85.47%
			29 671	509 149	22.61%
Greta Thunberg		GretaThunberg*	9 556	10 988	22.82%

Notes: Current index constitution can be found at https://www.boerse-frankfurt.de/indices/dax/constituents. * denotes accounts that are verified by Twitter. We merge all Merck tweets under @Merck and all Continental ones under @Conti_Press. % translated denotes the percentage of # tweets that was not originally made in English.

Figure 1: Environment vs. Non-environment Tweets by Entity, 2013-2023



firms to inform them of their environmental concerns and action. This might impact stock pricing through first- and second-round effects: the former if these same stakeholders are making investment decisions, and the latter if investing stakeholders value the wider public's perception of the firm. The intensity of a firm's climate communication might, thus, by itself have a positive effect on pricing. Alternatively, the public might find the affirming of climate concerns by corporations (especially pollutant ones) to be insincere and motivated purely by PR concerns, in which case this communication could have a negative price effect.

Greta Thunberg's communication can also have an impact on stock pricing if her frequently anti-capitalist rhetoric can negatively prime stakeholders and the public towards firms, and thus induce a negative price effect. Finally, the combination of Thunberg's and firms' communication can also be relevant: firstly, a tweet or series of tweets by Greta might put firms on the spotlight, enabling or amplifying the previously-discussed firm communication-stock pricing link. Additionally, it might provide firms with the chance to align with her views, rewarding those who do and/or punishing those that don't. Generally, we are looking at reputational gains and losses from corporate communication and their translation into stock pricing – in what follows, we design a strategy that allows for the evaluation of these possibilities.

Regarding the relationship between a firm's tweeting and their stock valuation, one must also be mindful of potential issues such as endogeneity and reverse causality. Endogeneity might be an issue if there is a common factor driving both corporate tweeting and prices, such as a natural disaster that directly impacts corporate activity. Reverse causality, on the other hand, can also be a problem as it is possible that the relationship between corporate tweeting and stock prices is bidirectional – while prices might respond to a firm's public stance on the environment, firms themselves might adjust their communication according to stock price movements. As the next section will show, we attempt to ease these concerns in several ways, namely using a metric of sustained, rather than immediate "greenness"; relating corporate communication to future price movements; leveraging weekend financial market closures and lowered corporate activity; and including relevant controls.

4.1 Design choices

From our original 2013-2023 daily data, we construct an over-the-weekend dataset where each observation corresponds to a given firm/weekend pair. Each Friday, firms are classified as being climate-focused or climate-neutral in their communication, according to the intensity of their climate-related communication over a rolling window. We

then use the fact that markets close over the weekend to evaluate the change in stock prices over this communication split and a "Greta treatment": the intensity of Greta Thunberg's own climate communication over the weekend (in estimations, we employ two different intensity thresholds for this variable). Leveraging the market closure in this way allows for an experiment-like framework where we expose these two firm groups, drawn from a pool of similar individuals (the DAX40 index), to the Greta treatment.

By determining the green status of firms prior to each weekend and using periods of market closure and lowered corporate activity, we curb the potential for reverse causality and are able to observe a cleaner linkage between Greta's communication and the prices of firms in each group. Obviously, firms can still be aware of a potential benefit or harm of participating in the climate discussion in general, but they will not be able to predict the timing or intensity of Greta's communication over the next weekend, at least severely limiting the likelihood of corporate communication as a direct response to Greta's behavior. Ex ante, we expect climate-focused firms to be more exposed to a Greta effect than climate-neutral ones, but they are likely to both be impacted as they directly compete for investment.

In trying to identify this Greta effect, we might also face endogeneity: it is possible that a third force is driving both corporate stock prices and Greta Thunberg's tweeting behavior, namely severe or numerous climate-related events and natural disasters happening throughout the world. To address this, we implement a currently-occurring natural events control, which we obtain from the EM-DAT database.[29] This dataset documents, for each instance of an environmental event, start and end dates, impacted countries and regions, and estimated impacts in terms of the number of affected people and the monetary value of damages. We consider all events above median severity, as measured through the cost of impacts or the number of affected people, as these are more likely to influence international debate; and exclude events in Germany to reduce the likelihood that they impact stock valuation through more direct avenues (e.g. destruction of operational or production infrastructure), in line with the findings of Bourdeau-Brien and Kryzanowski (2017). Table A2, in the appendix, breaks down all considered events by type.

Finally, given market closures, investor reactions to any events that might influence their decisions will be reflected only on Monday when the market reopens. We expect this to facilitate the detection of a Greta effect, as any change in prices it might cause will come into effect at once on Monday, when the market opens, sparing us the issue of having to define across what timespan a tweet made by Greta might impact stock prices. It also limits simultaneous firm activity or near-immediate responses that could otherwise blur our identification.

4.2 Regression framework

We employ a simple fixed effects baseline as follows:

$$\frac{Monday\,open_{i,t} - Friday\,close_{i,t}}{Friday\,close_{i,t}} = \alpha + \beta_1 Greta\,treatment_t^{\theta} + \beta_2 Green_{i,t} + \beta_3 Greta\,treatment_t^{\theta} \times Green_{i,t} + \beta_4 Env.\,events_t + \beta_5 WE\,corp.\,env.\,tweets_{i,t} + \beta_6 Friday\,close_{i,t} + Month\,FE + Year\,FE + Firm\,FE + \epsilon_{i,t} \quad (1)$$

Where $Monday\,open_{i,t}$ stands for the first price company i's stock is traded for on the Monday immediately after weekend t; $Friday\,close_{i,t}$ stands for the last price company i's stock was traded for on the Friday immediately preceding weekend t; $Greta\,treatment_t^{\theta}$ is a dummy variable equal to 1 if Greta Thunberg tweeted about the environment at least θ times during weekend t and 0 otherwise; $Green_{i,t}$ is a dummy variable equal to 1 if company i is classified as climate positive in the period immediately prior to weekend t and 0 otherwise; $Env.\,events_t$ is a control that stands for the total number of severe natural events and disasters currently occurring in the Friday immediately preceding weekend t; $WE\,corp.\,env.\,tweets_{i,t}$ stands for the number of tweets about the environment made by firm i during weekend t; $Month\,FE$, $Year\,FE$, and $Firm\,FE$ stand for month-, year, and firm-level fixed

[29] Which can be found at https://www.emdat.be/ and lists natural environmental events per day.

Table 2: Descriptive statistics

Full timeline (1/1/2013-24/04/2023)					
Variable	# Obs.	Mean	Std. Dev.	Min	Max
$\frac{Monday\ open_{i,t} - Friday\ close_{i,t}}{Friday\ close_{i,t}}$	17 380	0.0005	0.0121	0.1994	0.4946
$Greta\ treatment_t^{2+}$	17 695	0.3854	0.4867	0	1
$Greta\ treatment_t^{3+}$	17 695	0.3415	0.4742	0	1
$Green_{i,t}^{week,\ 1+}$	17 695	0.4268	0.4946	0	1
$Green_{i,t}^{week,\ 2+}$	17 695	0.2765	0.4473	0	1
$Green_{i,t}^{week,\ 3+}$	17 695	0.1832	0.3869	0	1
$Green_{i,t}^{week,\ 4+}$	17 695	0.1257	0.3316	0	1
$Green_{i,t}^{2week,\ 1+}$	17 695	0.5456	0.4979	0	1
$Green_{i,t}^{2week,\ 2+}$	17 695	0.4118	0.4922	0	1
$Green_{i,t}^{2week,\ 3+}$	17 695	0.3247	0.4683	0	1
$Green_{i,t}^{2week,\ 4+}$	17 695	0.2612	0.4393	0	1
$Environment\ events_t$	17 695	28.1812	9.3857	8	53
$WE\ corp.\ env.\ tweets_{i,t}$	17 695	0.0692	0.5338	0	42
$Friday\ Close_{i,t}$	17 380	66.8999	59.0073	2.7509	587.0031
Greta timeline (18/06/2018-24/04/2023)					
Variable	# Obs.	Mean	Std. Dev.	Min	Max
$\frac{Monday\ Open_{i,t} - Friday\ Close_{i,t}}{Friday\ Close_{i,t}}$	8 514	-0.0001	0.0145	-0.1994	0.4946
$Greta\ treatment_t^{2+}$	8 695	0.7802	0.4141	0	1
$Greta\ treatment_t^{3+}$	8 695	0.6910	0.4621	0	1
$Green_{i,t}^{week,\ 1+}$	8 695	0.5305	0.4991	0	1
$Green_{i,t}^{week,\ 2+}$	8 695	0.3568	0.4791	0	1
$Green_{i,t}^{week,\ 3+}$	8 695	0.2342	0.4235	0	1
$Green_{i,t}^{week,\ 4+}$	8 695	0.1577	0.3645	0	1
$Green_{i,t}^{2week,\ 1+}$	8 695	0.6578	0.4745	0	1
$Green_{i,t}^{2week,\ 2+}$	8 695	0.5208	0.4996	0	1
$Green_{i,t}^{2week,\ 3+}$	8 695	0.4199	0.4936	0	1
$Green_{i,t}^{2week,\ 4+}$	8 695	0.3417	0.4743	0	1
$Environment\ events_t$	8 695	31.4616	9.6511	16	53
$WE\ corp.\ env.\ tweets_{i,t}$	8 695	0.0840	0.6258	0	42
$Friday\ Close_{i,t}$	8 514	81.2406	73.5549	2.7509	587.0031

Notes: values between -0.0001 and 0.0001 are rounded down or up to those values.

effects; and $\epsilon_{i,t}$ is an error term. Table 2 presents descriptive statistics for all used variables in all forms they appear in. Table A3, in the appendix, presents their correlation matrices. We run regressions for two separate timeframes: 2018-2023, corresponding to the period when Greta Thunberg was active on Twitter, and 2013-2023, corresponding to our entire dataset.

5 Results and discussion

One of the world's leading companies for renewable energies and carbon neutral by 2040. (...)[30]

RWE AG (@RWE_AG), Twitter biography

[30]Source: https://twitter.com/RWE_AG.

5.1 Descriptive analysis

We begin by performing a descriptive analysis of the phenomena under study – stock prices, corporate tweeting and Thunberg's tweeting – across time. Figure 2 plots monthly environmental and non-environmental tweet volumes, comparing aggregate corporate values to Greta Thunberg's. We find that corporate environment tweets somewhat track Greta Thunberg's tweeting, at least until 2021 (which coincides with the period in which she was most active on the platform) – an indicator that firms might be paying attention to the environmental agenda in general (and Thunberg in particular) and aim to participate in it. While this could occasionally be said also of other corporate tweets, it does not appear to be a regularity in general. While this observation suggests that firms might look to Thunberg for inspiration when communicating through Twitter, it is by no means evidence of this as confounding factors are likely to be at play – for example, in case both firms and Thunberg are reacting to environmentally-relevant occurrences.

Figure 3, on the other hand, presents weekly aggregate values for DAX open prices and tweeting behavior – a very broad overlook of this potential relationship. At least until 2022, it seems to suggest a positive correlation between asset pricing and corporate environmental tweeting, with the 2019-20 and 2021-22 pricing humps coinciding with periods of especially intense corporate climate mentions. A fundamental change in this relationship seems to happen from the end of 2022 onward, as Thunberg becomes progressively less active. This could also be related to the definitive end of the COVID-19 era and begs further attention.

5.2 Estimation results

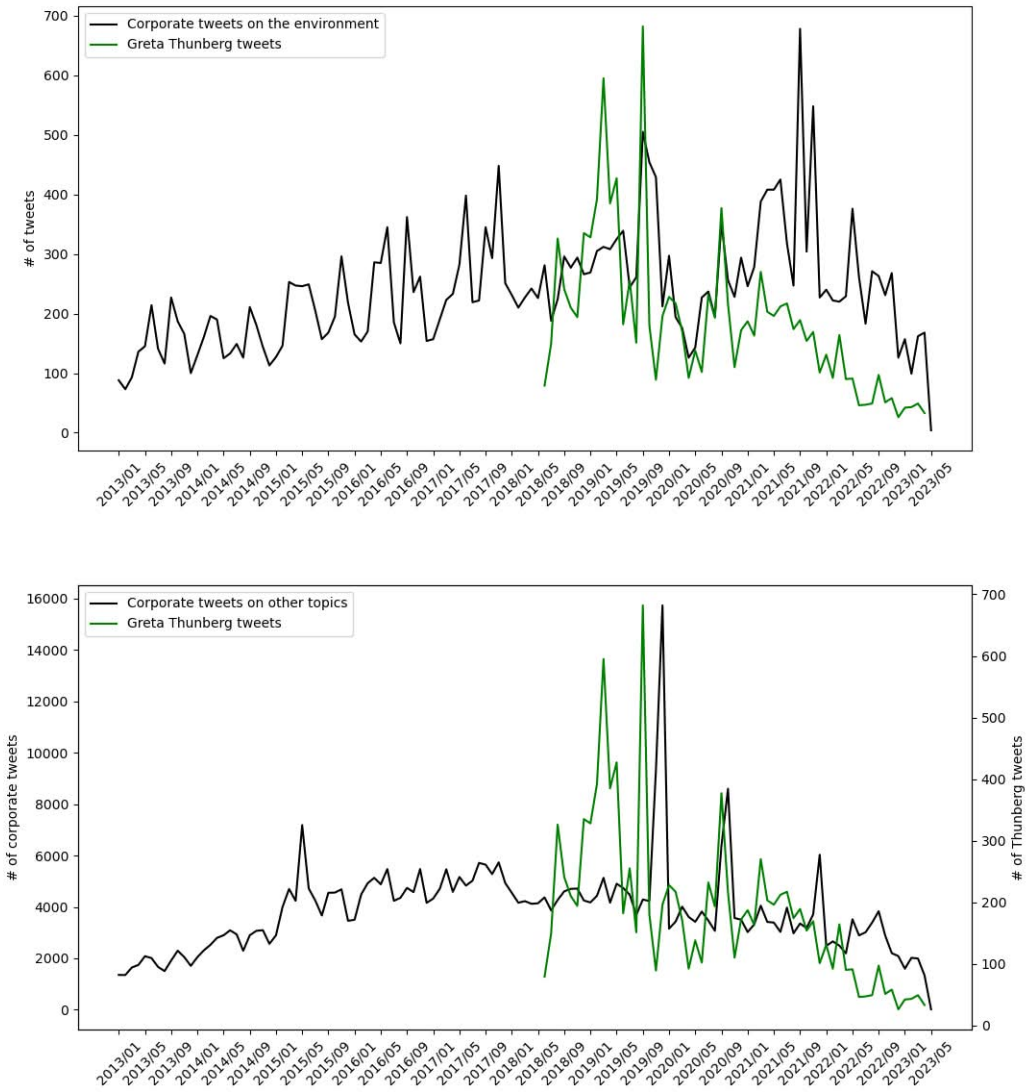
This section describes and discusses results. We begin with a simple formulation without interactions, presented in table 3. From these initial results, we can see that corporate climate communication does not seem to have a significant impact on stock price growth over the weekend. In the full timeframe, $Greta\,treatment_t^{\theta=2}$ can be seen as an imperfect dummy variable, being always equal to zero before June 2018 and often equal to one from then on (specifically, approximately 78% of the time). Thus, the fact that this variable is only found to be significant in the full timeframe suggests that, by herself, Greta Thunberg might have influenced DAX40 stock pricing negatively by emerging as a figurehead for the climate movement more than through individual communication efforts. That this impact is negative is an interesting finding that can suggest that her presence has made investors wary of large corporations, regardless of their efforts in climate-related communication.

It is further noteworthy that $Env.events_t$ are found to be significant and have a negative impact on the dependent variable. Together with the previous finding, a picture of a climate-aware investor that incorporates the action of public opinion leaders and natural phenomena in their trading decisions emerges as a plausible construct.

The previous results, however, do not speak to the potential interactions between corporate communication and Greta's. To investigate this topic, table 4 depicts our baseline estimation. These results are presented for $\theta = 2$ and $\theta = 3$ (which are equal to 1 in 78% and 69% of the Greta timeframe, respectively), using thresholds for $Green_{i,t}$ of at least one or at least two climate-related tweets over the previous week (equal to 1 in 53% and 36%, respectively, of the dataset in this timeframe). Table A4, in the appendix, further shows results for $Green_{i,t}$ thresholds of 3 and 4.

By including the interaction between $Greta\,treatment_t$ and $Green_{i,t}$, we find that it has a positive and significant associated coefficient across the board. This suggests that firms who participate in climate communication will be better off than those who don't, in terms of stock price growth over the weekend, when Greta Thunberg makes a public statement about the climate. The previous finding that the number of currently-occurring environmental events impacts stock prices negatively also holds across the board (albeit occasionally only at a 10% level). A further noteworthy finding is that the coefficients for $Greta\,treatment_t$ become substantially more negative when we increase its threshold, meaning that more intense Greta communication translates into lower stock price growth

Figure 2: Corporate and Thunberg tweet volumes per month



Notes: The top graph plots the volume of corporate tweets classified as referring to the environment against total Greta Thunberg tweet volume. The bottom graph does the same with other-topic corporate tweet volume.

Figure 3: DAX pricing and tweets about the environment, weekly

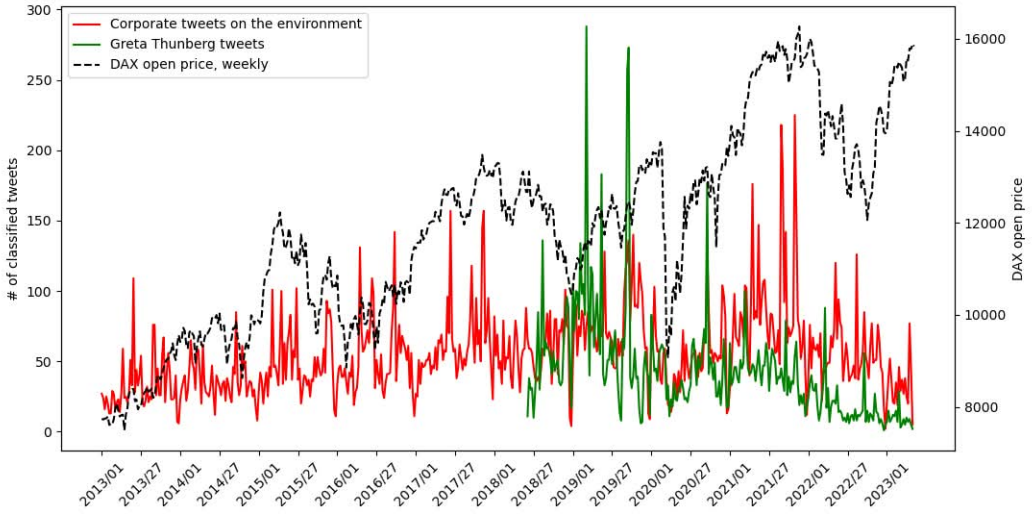


Table 3: Results: Initial (green firms: 1+ tweets over the week; Greta treatment: 2+ tweets over the weekend)

$Y = WE\,price\,gr$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Greta\,treatment_t$	-0.080***	-0.022	-0.074**	-0.021	-0.074**	-0.021	-0.073**	-0.020
	(0.028)	(0.038)	(0.028)	(0.038)	(0.028)	(0.038)	(0.028)	(0.038)
$Green_{i,t}$	-0.002	-0.038	-0.001	-0.037	-0.002	-0.037	-0.004	-0.037
	(0.027)	(0.051)	(0.027)	(0.051)	(0.027)	(0.051)	(0.027)	(0.052)
$Env.\,events_t$			-0.004***	-0.004**	-0.004***	-0.004**	-0.004***	-0.004**
			(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
$WE\,corp.\,env.\,tweets_t$					0.001	-0.010	0.002	-0.010
					(0.014)	(0.013)	(0.014)	(0.013)
$Friday\,close\,price_t$							-0.001**	-0.001
							(0.000)	(0.000)
Sample	2013-2023	2018-2023	2013-2023	2018-2023	2013-2023	2018-2023	2013-2023	2018-2023
Total # observations	17,380	8,514	17,380	8,514	17,380	8,514	17,380	8,514
R^2	0.025	0.034	0.025	0.034	0.025	0.034	0.026	0.035
Total # firms	38	38	38	38	38	38	38	38

Notes: Initial results using the variables of interest without interactions. Used samples alternate between the entire timeframe (January 7th 2013 - April 24th 2023) and the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has tweeted at least once about the environment in the 7 days leading to the respective Friday and 0 otherwise. Results are equivalent if the threshold for $Green_{i,t} = 1$ is increased to 2, 3, or 4 tweets about the environment in the preceding week. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table 4: Results: Baseline (green firms: # of tweets over the week)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.130***	-0.137**	-0.203***	-0.214***
	(0.048)	(0.059)	(0.044)	(0.047)
$Green_{i,t}$	-0.197**	-0.316***	-0.134*	-0.233***
	(0.094)	(0.084)	(0.075)	(0.062)
$Greta\,treatment_t \times Green_{i,t}$	0.208**	0.326***	0.145**	0.247***
	(0.081)	(0.096)	(0.065)	(0.076)
$Env.\,events_t$	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\,corp\,env.\,tweets_{i,t}$	-0.010	-0.010	-0.010	-0.009
	(0.013)	(0.013)	(0.013)	(0.013)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 1 week	2+, 1 week	1+, 1 week	2+, 1 week
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.035	0.037	0.036	0.037
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least one or two tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

over the weekend. This impact roughly cancels out the positive coefficient of the interaction term, implying that firms that do not tweet about the environment stand to lose from these doses of Greta tweets, while those who do tweet simply mitigate this effect. Finally, $Green_{i,t}$ by itself displays negative and significant coefficients for thresholds of at least one or two tweets over the last week, offering some evidence of a general negative effect of participating in the climate discussion on the dependent variable.

Overall, these results paint a complex image: there is strong evidence that, if firms wish to maximize stock price growth over the weekend, they should participate in the climate discussion when Greta Thunberg does, but not in general. Additionally, it is not so much that firms who behave in this manner will benefit – rather, it is those who don't who will be harmed. One possible interpretation for these results is that the general consensus regarding corporate climate communication sees it more as greenwashing than genuine concern, and that Thunberg, given her often adversarial speech towards corporations, might represent and foster this feeling. Given tweeting simultaneously with her counteracts this effect, one might suggest markets see alignment with Thunberg as a sign of legitimate commitment towards sustainability, or at least as a good business practice.

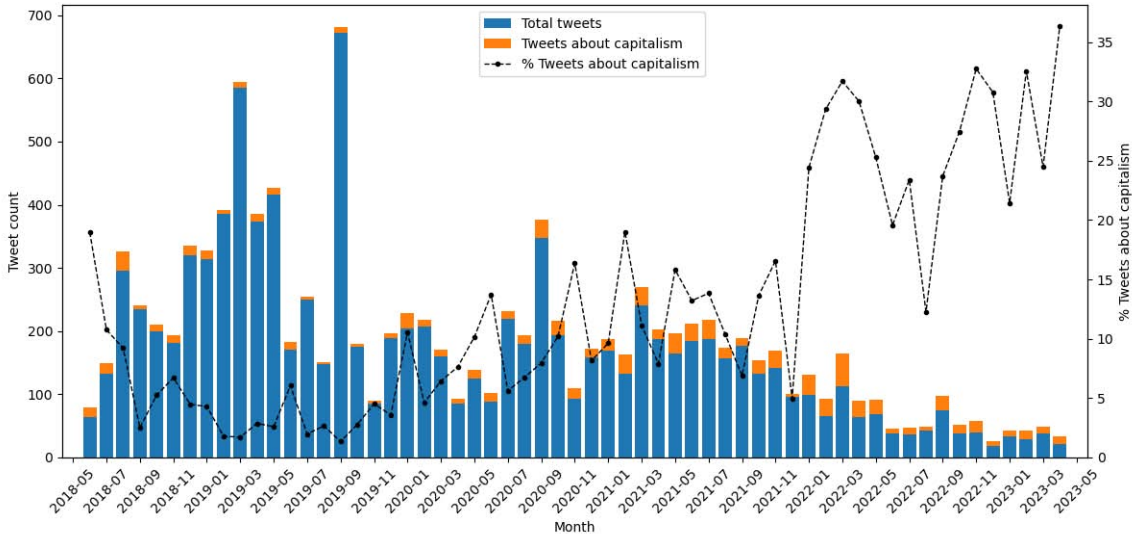
Finally, table 5 presents the full baseline estimation for the full, 2013-2023 timeframe. Through this broader scope, the interaction effects disappear and we just find the negative (and highly significant) effect of $Greta\,treatment_t$ on weekend price growth. This seems to back our initial observation that Greta's emergence as a paladin of climate topics had a negative overall effect on DAX40 stock prices; but also suggests that the intensity of her communication matters as $Greta\,treatment_t^{\theta=3}$ presents coefficients that are generally twice as large as $Greta\,treatment_t^{\theta=2}$. Once more, table A5 presents results for $Green_{i,t}$ thresholds of 3 and 4. That the results for the interaction term disappear using this broader scope does not mean they are not present or substantial – rather, that they are drowned out by the "appearance of Greta" effect.

Table 5: Results: Baseline, full timeframe (green firms: # of tweets over the week)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.079**	-0.091***	-0.165***	-0.180***
	(0.030)	(0.028)	(0.029)	(0.027)
$Green_{i,t}$	-0.009	-0.051	-0.005	-0.046
	(0.032)	(0.032)	(0.030)	(0.027)
$Greta\,treatment_t \times Green_{i,t}$	0.012	0.053	0.004	0.050
	(0.038)	(0.043)	(0.038)	(0.039)
$Env.\,events_t$	-0.004***	-0.004***	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	0.002	0.003	0.002	0.003
	(0.014)	(0.014)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 1 week	2+, 1 week	1+, 1 week	2+, 1 week
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2013-2023	2013-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	17,380	17,380	17,380	17,380
R^2	0.026	0.026	0.027	0.027
Total # firms	38	38	38	38

Notes: The sample consists of our full dataset (January 7th 2013 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least one or two tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10%, respectively. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Figure 4: Thunberg tweets about capitalism, 2018-2023



5.3 Additional results

This section presents and discusses results that expand the baseline's scope based on tweet sentiment and content. We have so far focused on Thunberg's communication as uni-dimensional – a voice for the environment. However, as discussed above multiple times, she also frequently engages with corporations and economic agents adversarially. As such, in what follows we firstly quantify the degree to which Greta directly (and negatively) addresses capitalism and employ it in estimations as an alternative Greta treatment.

To obtain this variable, we employ a GPT classifier on all Greta Thunberg tweets in our dataset.[31] Out of her 10 998 tweets, 961 are identified as being about capitalism – approximately 9%. Figure 4 graphically depicts this series, showing firstly how this brand of communication present but relatively uncommon, and secondly how it became a more prominent part of her message towards the end of our timeframe – it seems that, by mid-2023, Greta had maintained or even slightly intensified her outward communication against corporations, but in general communicated much less than before. Table 6 presents our baseline regression with an alternative Greta treatment, equal to 1 if she made at least one anti-capitalism tweet, for specifications (1) and (2), or at least two for (3) and (4). These results are generally in line with previous ones, which could mean that the content of Thunberg's communication is not as important, in this regard, as its simple occurrence.

As a second exercise in deeper exploration of the possibilities Twitter data offer, we look at Greta and corporate tweet sentiment and emotion to evaluate whether alignment in this aspect is of any relevance. To begin, we apply the NRC Word-Emotion Association Lexicon[32] (Mohammad and Turney, 2013) to all tweets through the NRCLex Python package. This classifier makes use of a pre-labelled dataset to classify each document along an axis of eight emotions and two sentiments, the former being anger, fear, anticipation, trust, surprise, sadness, joy, and disgust; and the latter simply being positive and negative. Figures 5 and 6 illustrates this task: figure 5 presents the sentiment split for firms and Greta, showing how most of corporate communication is apparently negative, while Greta achieves a more balanced split (which she kept relatively constant across time). Figure 6, on the other hand, shows all emotions and sentiments in a stacked bar per individual. Note that each tweet can be flagged for multiple

[31] GPT 4.1, max_tokens=1, temperature=0, prompt= "You are a research assistant classifying tweets by public figures. Decide whether the following tweet refers to capitalism or corporate activity in a negative way. Negative tweets might blame specific firms or generic profit-seeking activity for negative phenomena; call people or governments to action against them; or put into question their good intentions. Label each tweet with either 1 (the tweet refers to capitalism or corporations negatively) or 0 (the tweet does not refer to capitalism or corporations negatively). Please begin your answer with your chosen label (1 or 0). Tweet:{text}".

[32] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Table 6: Results: Anti-capitalism communication

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.186***	-0.172***	-0.278***	-0.266***
	(0.029)	(0.033)	(0.046)	(0.039)
$Green_{i,t}$	-0.099*	-0.140***	-0.073	-0.113***
	(0.058)	(0.045)	(0.053)	(0.040)
$Greta\,treatment_t \times Green_{i,t}$	0.147***	0.177***	0.223***	0.289***
	(0.045)	(0.054)	(0.073)	(0.069)
$Env.\,events_t$	-0.003*	-0.003**	-0.004**	-0.004**
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\,corp\,env.\,tweets_{i,t}$	-0.010	-0.009	-0.010	-0.008
	(0.013)	(0.013)	(0.013)	(0.013)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 1 week	2+, 1 week	1+, 1 week	2+, 1 week
$Greta\,treatment_t$	1+, anti-capitalism	1+, anti-capitalism	2+, anti-capitalism	2+, anti-capitalism
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.036	0.037	0.037	0.038
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least once, in specifications (1) and (2), or twice, for (3) and (4), about capitalism in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least one or two tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Figure 5: Corporate and Thunberg tweets by sentiment

Notes: The top graph shows the positive/negative overall sentiment split in the dataset by individual. The bottom graph details Greta Thunberg's split across time.

(even all) emotions and sentiments, making the total counts much larger than the total number of tweets.

Table 7 presents results for this study. In general, we replace $Green_{i,t}$ with $Aligned_{i,t}$ – for specifications (1) and (2), this variable takes the value of one if both the firm in question (over the last week) and and Greta (over the weekend) share their average sentiment (positive or negative), and zero otherwise. For specifications (3) to (6), we focus on emotions instead. We first compute an alignment counter that, for each firm, counts the number of emotions that both their tweets (over the week) and Greta's (over the weekend) displayed. $Aligned_{i,t}$ then takes the value of 1 if this value is above the sample-wide median (5 or more, 56% of the sample) or 75th percentile (7 or more, 27%), depending on the specification.

Results suggest that corporate alignment with Thunberg on sentiment is not consequential in this framework. Alignment on emotions, on the other hand, seems to relate positively with the dependent variable when $Greta\,treatment = 1$, which might indicate that there is more to the effect we have so far found than simple message content – delivery and tone, if consistent with Greta's, apparently can also attenuate the negative direct $Greta\,treatment$ effect.

Figure 6: DAX pricing and tweets about the environment, weekly

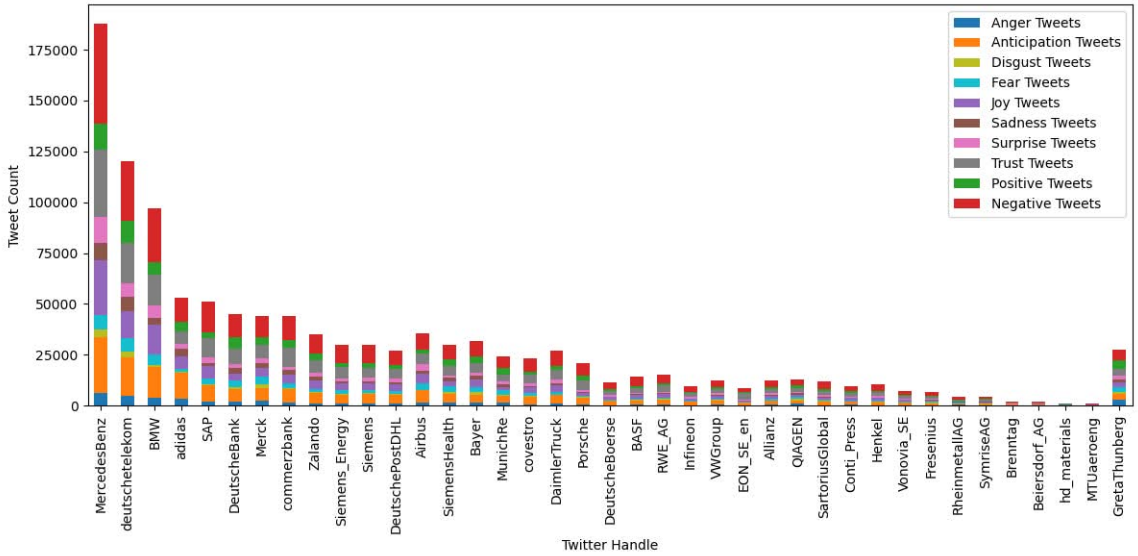


Table 7: Results: Sentiment and emotion alignment

$Y = \frac{Monday\ Open_{i,t} - Friday\ Close_{i,t}}{Friday\ Close_{i,t}}$	(1)	(2)	(3)	(4)	(5)	(6)
$Greta\ treatment_t$	0.023	-0.080	-0.084	-0.045	-0.216***	-0.159***
	(0.055)	(0.048)	(0.054)	(0.044)	(0.061)	(0.038)
$Aligned_{i,t}$	-0.025	-0.027	-0.131*	-0.128	-0.110*	-0.127
	(0.076)	(0.062)	(0.066)	(0.096)	(0.059)	(0.087)
$Greta\ treatment_t \times Aligned_{i,t}$	-0.098	-0.100	0.174**	0.173*	0.201**	0.203**
	(0.081)	(0.073)	(0.084)	(0.102)	(0.090)	(0.094)
$Env.\ events_t$	-0.004**	-0.003	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$WE\ corp.\ env.\ tweets_t$	-0.011	-0.011	-0.012	-0.011	-0.012	-0.011
	(0.013)	(0.014)	(0.013)	(0.013)	(0.014)	(0.013)
$Friday\ close\ price_t$	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$Aligned_{i,t}$	Avg sent.	Avg sent.	Emo > med.	Emo > 75p.	Emo > med.	Emo > 75p.
$Greta\ treatment_t$	2+, weekend	3+, weekend	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514	8,514	8,514
R^2	0.036	0.037	0.035	0.035	0.036	0.036
Total # firms	38	38	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table 8: Results: Baseline (green firms: # of tweets over the fortnight)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.185***	-0.107**	-0.221***	-0.167***
	(0.054)	(0.048)	(0.041)	(0.039)
$Green_{i,t}$	-0.297***	-0.186*	-0.203***	-0.117
	(0.080)	(0.100)	(0.058)	(0.079)
$Greta\,treatment_t \times Green_{i,t}$	0.246***	0.159*	0.145***	0.081
	(0.073)	(0.090)	(0.052)	(0.064)
$Env.\,events_t$	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\,corp\,env.\,tweets_{i,t}$	-0.010	-0.010	-0.009	-0.009
	(0.013)	(0.013)	(0.013)	(0.013)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 2 weeks	2+, 2 weeks	1+, 2 weeks	2+, 2 weeks
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.036	0.035	0.036	0.036
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least one or two tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

6 Robustness

> *We hope you have a pleasant journey with us, as we share the latest news and views on our family of aircraft, sustainable aviation and much more.* (...)[33]
>
> Airbus SE (@Airbus), Twitter biography

This section explores alternative formulations that lend robustness to our previous analysis, namely evaluating the "greenness" of firms over a 2-week period rather than a single week and using a static green measure based on three different measures: firm ESG scores, the level of pollutive emissions associated with their sector of activity and their overall climate communication.

6.1 Green fortnights

Table 8 describes our baseline estimation measuring green scores over a fortnight, rather than just a week; while table 9 does the same for the 2013-2023 period rather than just the Greta timeframe. Results are very similar and yield equivalent conclusions. As before, tables A6 and A7 present these results when using $Green_{i,t}$ thresholds of 3 and 4.

[33]Source: https://twitter.com/Airbus.

Table 9: Results: Baseline, full timeframe (green firms: # of tweets over the fortnight)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.079**	-0.091***	-0.165***	-0.180***
	(0.030)	(0.028)	(0.029)	(0.027)
$Green_{i,t}$	-0.009	-0.051	-0.005	-0.046
	(0.032)	(0.032)	(0.030)	(0.027)
$Greta\,treatment_t \times Green_{i,t}$	0.012	0.053	0.004	0.050
	(0.038)	(0.043)	(0.038)	(0.039)
$Env.\,events_t$	-0.004***	-0.004***	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	0.002	0.003	0.002	0.003
	(0.014)	(0.014)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 2 weeks	2+, 2 weeks	1+, 2 weeks	2+, 2 weeks
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2013-2023	2013-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	17,380	17,380	17,380	17,380
R^2	0.026	0.026	0.027	0.027
Total # firms	38	38	38	38

Notes: The sample consists of our full dataset (January 7th 2013 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least one or two tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table 10: Results: Static $Green_{i,t}$ (green firms: above-mean total # of tweets about the environment)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.043	-0.129***	-0.064**	-0.151***
	(0.059)	(0.046)	(0.029)	(0.026)
$Greta\,treatment_t \times Green_{i,t}$	0.055	0.012	-0.022	-0.031
	(0.066)	(0.052)	(0.026)	(0.025)
$Env.\,events_t$	-0.004**	-0.003*	-0.004***	-0.002*
	(0.002)	(0.002)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	-0.011	-0.010	0.001	0.002
	(0.013)	(0.013)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	Above-mean	Above-mean	Above-mean	Above-mean
$Greta\,treatment_t$	2+, weekend	3+, weekend	2+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	17,380	17,380
R^2	0.034	0.035	0.026	0.027
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023) in specifications (1) and (2) and the full dataset (January 7th 2013 - April 24th 2023) in specification (3) and (4). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (3), or three times, for (2) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made an above-mean number of tweets about the environment in the period under analysis and 0 otherwise (this variable is omitted, save for the interaction term, due to the inclusion of firm-level fixed effects). Constructing this metric with the median value yields equivalent results. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

6.2 Static greenness

In our baseline estimations, we flexibly classify firms over the climate-positive/climate-neutral axis using their recent communication. One might argue, however, that the public might not be as minute in their analysis, and thus that we could better capture the public's perception of corporate greenness through a persistent indicator. To address these concerns, we re-run our baseline estimation using a static measure of climate-positive communication. Specifically, we classify a firm as being green ($Green_{i,t} = 1$) if it has made an above-mean amount of tweets about the environment in the period under analysis – table 10 reports these results.

We can observe that this static $Green_{i,t}$ is not found to significantly impact the dependent variable – just having tweeted more about the environment in general does not interact in any noticeable way with $Greta\,treatment_t$. Apparently, the timing of corporate communication on these topics is relevant and, to reap the benefits of alignment with opinion leaders, engaging in their discussions contemporaneously is important.

6.3 ESG-based greenness

There is also the possibility that investors are more concerned with corporate activity than communication when evaluating the climate positions of firms, and that climate communication is either secondary to or a proxy of their ESG compliance. To evaluate whether these factors might be behind our results, we turn to the DAX 50 ESG index, which lists the 50 largest stocks in the HDAX universe after an ESG-compliance based selection process. Given this index was only introduced in March 2020, we cannot allow for composition changes before that period. As such, we propose two different applications: a static one, where $Green_{i,t}$ is equal to 1 if the firm in question was included in the first version of the DAX 50 ESG; and a flexible one where we allow for firm entry and exit.

Table 11: Results: Static $Green_{i,t}$ (green firms: included in DAX 50 ESG)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.071	-0.149	-0.044	-0.130***
	(0.119)	(0.090)	(0.033)	(0.032)
$Greta\,treatment_t \times Green_i$	0.070	0.035	-0.040	-0.046
	(0.123)	(0.095)	(0.028)	(0.027)
$Env.\,events_t$	-0.004**	-0.003*	-0.004***	-0.002*
	(0.002)	(0.002)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	-0.011	-0.010	0.002	0.003
	(0.013)	(0.013)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	DAX 50 ESG	DAX 50 ESG	DAX 50 ESG	DAX 50 ESG
$Greta\,treatment_t$	2+, weekend	3+, weekend	2+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	17,380	17,380
R^2	0.034	0.035	0.026	0.027
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023) in specifications (1) and (2) and the full dataset (January 7th 2013 - April 24th 2023) in specification (3) and (4). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (3), or three times, for (2) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 for firms included in the initial, March 2020 composition of the DAX 50 ESG index and 0 otherwise (this variable is omitted, save for the interaction term, due to the inclusion of firm-level fixed effects). Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Tables 11 and 12, respectively, present results for the static and flexible applications. They broadly show no impact of DAX 50 ESG membership on stock prices, and do not suggest that this factor influences firm susceptibility or exposure to Greta Thunberg's communication.

6.4 Emissions-based greenness

One could also argue that the market (and the public) might be more concerned about pollution than a broader ESG evaluation. To test this hypothesis, we turn to Bechtel et al. (2019), who, for use in their analysis, compute industry-level pollution measures for several countries, including Germany, drawing on different sources for 2011 data. Their baseline formulation employs total greenhouse gas (GHG) emissions for each industry, measured in millions of tonnes of produced Co2-equivalent gases, which they use to split industries into less-pollutive if they are below the median observed value and more-pollutive otherwise.

For this subsection, we employ exactly this measure as $Green_i$, which yields the more/less-pollutive split shown in table 13. By examining it, we find that our sample is not particularly well-balanced over this split – only 23% of all observations are in the $Green_i = 1$ field, and only a few sectors are represented overall. Nonetheless, the objectiveness of this measure makes it valuable for our analysis – table 14 presents results using this variable.

With this formulation, a single specification hints at the significance of the $Greta\,treatment \times Green_i$ interaction – (3), using the full timeframe and the weaker Greta treatment. Note that this is the only formulation for which this result occurs – making the green classification less stringent, for example with $Green_i = 0$ only for sectors in the top emissions tercile or quartile, yields no significance for any specification.

While it is too weak and isolated to motivate robust conclusions, this result suggests that the appearance of Greta Thunberg as a figurehead of the environmentalist movement could have had a negative impact on the market

Table 12: Results: Flexible $Green_{i,t}$ (green firms: included in DAX 50 ESG)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.111	-0.182*	-0.184	-0.250**
	(0.127)	(0.095)	(0.159)	(0.114)
$Green_{i,t}$	-0.065	-0.022	-0.133	-0.081
	(0.162)	(0.143)	(0.195)	(0.174)
$Greta\,treatment_t \times Green_{i,t}$	0.120	0.078	0.129	0.082
	(0.133)	(0.104)	(0.164)	(0.125)
$Env.\,events_t$	-0.004**	-0.003*	-0.054***	-0.054***
	(0.002)	(0.002)	(0.004)	(0.004)
$WE\,corp\,env.\,tweets_{i,t}$	-0.011	-0.010	-0.015	-0.014
	(0.013)	(0.013)	(0.018)	(0.018)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.000
	(0.000)	(0.000)	(0.001)	(0.001)
$Green_{i,t}$	DAX 50 ESG	DAX 50 ESG	DAX 50 ESG	DAX 50 ESG
$Greta\,treatment_t$	2+, weekend	3+, weekend	2+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2020-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	5,539	5,539
R^2	0.035	0.036	0.066	0.067
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023) in specifications (1) and (2) and the DAX 50 ESG timeframe (March 2nd 2020 - April 24th 2023) in specification (3) and (4). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (3), or three times, for (2) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 for firms included in the DAX 50 ESG index at the time of each observation and 0 otherwise (for specifications (1) and (2), this variable takes the March 2020 values for all prior observations). Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table 13: Firms and sectoral emissions

Firm	Sector	ISIC	*Green*
Allianz SE	Insurance	Financial and insurance activities	1
Deutsche Bank AG	Banking	Financial and insurance activities	1
Deutsche Börse AG	Financial Services	Financial and insurance activities	1
Münchener Rück AG	Insurance	Financial and insurance activities	1
SAP SE	Software	Information and communication	1
Vonovia SE	Financial Services	Financial and insurance activities	1
Commerzbank AG	Banking	Financial and insurance activities	1
Deutsche Telekom AG	Telecommunication	Information and communication	1
Adidas AG	Consumer goods	Manufacturing	0
Airbus SE	Industrial	Manufacturing	0
BASF SE	Chemicals	Manufacturing	0
BMW AG St	Automobile	Manufacturing	0
Bayer AG	Chemicals	Manufacturing	0
Beiersdorf AG	Consumer goods	Manufacturing	0
Brenntag SE	Industrial	Manufacturing	0
Continental AG	Automobile	Manufacturing	0
Covestro AG	Chemicals	Manufacturing	0
Daimler Truck Holding AG	Industrial	Manufacturing	0
Deutsche Post AG	Transportation & Logistics	Transportation and storage	0
E.ON SE	Utilities	Electricity, gas, steam and air conditioning supply	0
Fresenius SE & Co. KGaA	Pharma & Healthcare	Manufacturing	0
Heidelberg Materials AG	Construction	Construction	0
Henkel AG & Co. KGaA Vz	Consumer goods	Manufacturing	0
Infineon Technologies AG	Technology	Manufacturing	0
MTU Aero Engines AG	Industrial	Manufacturing	0
Mercedes Benz Group AG	Automobile	Manufacturing	0
Merck KGaA	Pharma & Healthcare	Manufacturing	0
Porsche AG Vz	Automobile	Manufacturing	0
Qiagen N. V.	Pharma & Healthcare	Manufacturing	0
RWE AG St	Utilities	Electricity, gas, steam and air conditioning supply	0
Rheinmetall AG	Industrial	Manufacturing	0
Sartorius Ag Vz	Pharma & Healthcare	Manufacturing	0
Siemens AG	Industrial	Manufacturing	0
Siemens Healthineers AG	Pharma & Healthcare	Manufacturing	0
Siemens Energy AG	Industrial	Manufacturing	0
Symrise AG	Chemicals	Manufacturing	0
Volkswagen Group AG Vz	Automobile	Manufacturing	0
Zalando SE	Retail	Wholesale and retail trade; repair of motor vehicles	0

Notes: "Sector" is as obtained from each firm's profile on https://www.boerse-frankfurt.de/indices/dax/constituents. ISIC (International Standard Industrial Classification) is a UN industry classification chart; equivalences to ISIC categories were made by the authors based on Bechtel et al. (2019) and its supplementary materials. For more information, see https://unstats.un.org/unsd/classifications/Econ/isic.

Table 14: Results: Emissions-based (static) $Green_{i,t}$ (green firms: lower GHG emissions)

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.020	-0.126***	-0.062**	-0.154***
	(0.045)	(0.035)	(0.028)	(0.025)
$Greta\,treatment_t \times Green_i$	-0.003	0.008	-0.048*	-0.040
	(0.060)	(0.049)	(0.024)	(0.026)
$Env.\,events_t$	-0.004**	-0.003*	-0.004***	-0.002*
	(0.002)	(0.002)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	-0.011	-0.010	0.002	0.002
	(0.013)	(0.013)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	Low emissions	Low emissions	Low emissions	Low emissions
$Greta\,treatment_t$	2+, weekend	3+, weekend	2+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	17,380	17,380
R^2	0.034	0.035	0.026	0.027
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023) in specifications (1) and (2) and the full dataset (January 7th 2013 - April 24th 2023) in specification (3) and (4). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (3), or three times, for (2) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 for firms whose sectors were responsible for a below-median amount of total GHG emissions and 0 otherwise (this variable is omitted, save for the interaction term, due to the inclusion of firm-level fixed effects). Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

performance of cleaner industries, which would certainly be an interesting phenomenon to study if it were to be verified. Other than this, this version of $Green_i$ also seems to bear no impact on the dependent variable and does not significantly interact with Greta's communication in the 2018-2023 timeline. Together with the ESG-based analysis, this strongly suggests that our previous results are indeed driven by communicational factors, rather than sectoral or ESG evaluations by investors.

6.5 Miscellaneous

Finally, table 15 present additional robustness tests that did not fit into the previous sections. In specifications (1) and (2), $Greta\,treatment_{i,t}$ takes the form of a placebo: equal to 1 if Greta made no tweet about the environment over the last weekend (happens in approximately 10% of the 2018-2023 timeframe) and 0 otherwise. Specifications (3) and (4), on the other hand, make $Greta\,treatment_{i,t}$ continuous, employing the count of tweets made by her during the weekend.[34]

As expected and desirable, we find that the absence of a Greta treatment has no impact on the dependent variable. Specifications (3) and (4), on the other hand, further reinforce the idea that the intensity of Greta's communication matters: the more she tweets about the environment over the weekend, the larger her impacts on the dependent variable – more negative through the direct channel, and more positive through the interaction channel.

[34]Note that, untreated, this variable was exceptionally skewed – it ranged from 0 to 49, but its 75th and 90th percentiles, respectively, were at 5 and 11. We thus capped the variable at 10, under the assumption that any value above that is functionally the same: a weekend of intense Greta communication.

Table 15: Results: Placebo and continuous Greta treatment

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.032	-0.015	-0.015***	-0.014***
	(0.079)	(0.061)	(0.005)	(0.005)
$Green_{i,t}$	-0.028	-0.045	-0.110	-0.161**
	(0.054)	(0.038)	(0.071)	(0.068)
$Greta\,treatment_t \times Green_i$	-0.079	-0.156	0.014**	0.019**
	(0.116)	(0.103)	(0.006)	(0.009)
$Env.\,events_t$	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\,corp\,env.\,tweets_{i,t}$	-0.009	-0.008	-0.010	-0.009
	(0.012)	(0.012)	(0.013)	(0.013)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	1+, 1 week	2+, 1 week	1+, 1 week	2+, 1 week
$Greta\,treatment_t$	0, weekend	0, weekend	Cont., 0-10	Cont., 0-10
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.035	0.035	0.035	0.035
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023) in specifications (1) and (2) and the full dataset (January 7th 2013 - April 24th 2023) in specification (3) and (4). Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

7 Conclusion

> *At BASF, we create chemistry for a sustainable future. We combine economic success with environmental protection and social responsibility.*[35]
>
> BASF SE (@BASF), Twitter biography

Greta Thunberg has significantly affected contemporary discussion on climate change. In one of the first studies to evaluate her impact on corporate tweeting behavior and market valuation, we find that firms' engagement with the climate agenda, as measured through their tweeting behavior, has an impact on their stock market outcomes – namely, that communicating on climate topics contemporaneously with Thunberg significantly and positively impacts stock price growth for DAX40 firms. We also find some evidence that emotional alignment with her communication has an analogous effect.

Higher volumes of Greta and corporate tweets about the environment, however, are by themselves associated with lower stock price growth – Greta Thunberg, by spearheading a movement that is in many regards anti-corporations, is found to hamper their valuations when speaking out; and firms might be seen as guilty of greenwashing when communicating too frequently on climate topics. Our overall conclusion from these findings is that, for stock-price maximizing firms, climate communication should be made in a timely and parsimonious manner – the former so as to align with opinion leaders, and the latter to avoid accusations of insincerity. Our analysis suggests that these findings are not driven by firm operations-related considerations (e.g. meeting of ESG standards, greenhouse gas emissions or sector of activity).

By employing Twitter data to measure communication, we obtain excellent granularity, are able to leverage

[35]Source: https://twitter.com/BASF.

daily tweet output and stock movements, and benefit from real-life, rather than lab observation. To tackle potential endogeneity and reverse causality issues, we employ an over-the-weekend, experiment-like formulation that splits the largest German firms into climate-positive and climate-neutral through an analysis of their sustained, rather than immediate corporate communication. We then expose them to a Greta treatment – weekends of more intense Greta communication on climate topics – and evaluate the differential impact on each group. By leveraging weekend financial market closures and lowered corporate activity, we curb the potential for instantaneous corporate communication responses and disentangle the relationship from other external factors that might muddy a daily analysis.

This study contributes to the growing fields of knowledge on corporate twitter usage, the role of climate activists as opinion leaders, and how these two phenomena might interact. It provides relevant takeaways both for the industry, which should monitor (as they likely already do) such opinion leaders and adjust their communication accordingly; and for future research, which should look into the existence of other similar relationships. Expanding on our content analysis of corporate and activist communication is an especially interesting further research avenue – employing LLM-based classifiers for sentiment and emotion analysis, for example, is likely to allow for a more precise and descriptive set of labels which might power deeper conclusions and a better evaluation of whether it is the content or the simple existence of environmentally-minded communication that is relevant. Finally, whether the negative association between corporate climate-positiveness and stock price growth is is due to unfair stigmatization or actual dishonest corporate intentions also remains to be answered and could be the topic of subsequent studies.

References

[1] Amichai-Hamburger, Y., McKenna, K. Y. A., Tal, S.-A. (2008). "E-empowerment: Empowerment by the Internet." Computers in Human Behavior 24(5): 1776–1789.

[2] Banerjee, S. B. (2002). "Corporate environmentalism: the construct and its measurement." Journal of Business Research 55(3): 177–191.

[3] Barrie, C., Ho, J. C.-t. (2021). "academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint." Journal of Open Source Software 6(62): 3272.

[4] Barnes, N. G., Lescault, A. M. (2013). "LinkedIn Rules But Sales Potential May Lie with Twitter: The 2013 Inc. 500 and Social Media," Center for Marketing Research of the University of Massachusetts Dartmouth.

[5] Barnes, N. G., Mazzola, A., Killeen, M. (2020). "Oversaturation & disengagement: The 2019 fortune 500 social media dance." Center for Marketing Research of the University of Massachusetts Dartmouth.

[6] Bechtel, M. M., Genovese, F., Scheve, K. F. (2019). "Interests, Norms and Support for the Provision of Global Public Goods: The Case of Climate Co-operation." British Journal of Political Science 49(4): 1333-1355.

[7] Bourdeau-Brien, M., Kryzanowski, L. (2017). "The impact of natural disasters on the stock returns and volatilities of local firms." The Quarterly Review of Economics and Finance 63: 259–270.

[8] Bowen, F. E. (2000). "Environmental visibility: A trigger of green organizational response?" Corporate Environmental Responsibility 107: 279–94.

[9] Capriotti, P., Ruesja, L. (2018). "How CEOs use Twitter: A comparative analysis of Global and Latin American companies." International Journal of Information Management 39: 242–48.

[10] Castelo Branco, M., Lima Rodrigues, L. (2008). "Social responsibility disclosure: A study of proxies for the public visibility of Portuguese banks." British Accounting Review 40(2): 161–81.

[11] Cadez, S., Czerny, A., Letmathe, P. (2019). "Stakeholder pressures and corporate climate change mitigation strategies." Business Strategy and the Environment 28(1): 1–14.

[12] Cho, M., Furey, L. D., Mohr, T. (2016). "Communicating Corporate Social Responsibility on Social Media: Strategies, Stakeholders, and Public Engagement on Corporate Facebook." Business and Professional Communication Quarterly, 80(1), 52–69.

[13] Dahal, B., Kumar, S. A. P., Li, Z. (2019). "Topic modeling and sentiment analysis of global climate change tweets." Social Network Analysis and Mining 9, 24.

[14] Delmas, M. A., Toffel, M. W. (2004). "Stakeholders and Environmental Management Practices: An Institutional Framework." Business Strategy and the Environment 13(4): 209-222.

[15] Delmas, M. A., Toffel, M. W. (2012). "Institutional Pressures and Organizational Characteristics: Implications for Environmental Strategy." In P. Bansal & A. J. Hoffman (Eds.), The Oxford Handbook of Business and the Natural Environment.

[16] Dietsche, C., Lautermann, C., Westerman, U. (2019). "CSR-Reporting in Deutschland 2018. Ergebnisse, Trends Branchenauswertungen und eine Analyse der Berichterstattung über die SDGs im Ranking der Nachhaltigkeitsberichte." Institut Für ökologische Wirtschaftsforschung Undfuture e.V.

[17] Downie, J., Stubbs, W. (2012). "Corporate carbon strategies and greenhouse gas emission assessments: The implications of scope 3 emission factor selection." Business Strategy and the Environment 21(6): 412–22.

[18] Ebner, M.; Mühlburger, H.; Schaffert, S.; Schiefner, M.; Reinhardt, W., Wheeler, S. (2010). "Getting granular on Twitter: Tweets from a conference and their limited usefulness for non-participants." IFIP Advances in Information and Communication Technology 324: 102–13.

[19] Elgesem, D., Brüggemann, M. (2023). "Polarisation or just differences in opinion: How and why Facebook users disagree about Greta Thunberg." European Journal of Communication, 38(3): 237-254.

[20] European Commission (2011). "Corporate Social Responsibility: a new definition, a new agenda for action." Memo/11/730.

[21] European Commission (2017). Special Eurobarometer 459 Report: Climate Change.

[22] Fernandez Gago, R., Nieto Antolín, M. (2004) "Stakeholder salience in corporate environmental strategy." Corporate Governance 4(3): 65–76.

[23] Fieseler, C., Fleck, M., Meckel, M. (2010). "Corporate social responsibility in the blogosphere." Journal of Business Ethics 91(4): 599–614.

[24] Freeman, R. E. (1984). "Strategic management: A stakeholder theory." Journal of Management Studies 39(1): 1–21.

[25] Gomez-Carrasco, P., Michelon, G. (2017). "The Power of Stakeholders' Voice: The Effects of Social Media Activism on Stock Markets." Business Strategy and the Environment, 26(6): 855– 872.

[26] Greiwe, J., Schönbohm, A. (2011). "A KPI based study on the scope and quality of sustainability reporting by the DAX30 companies." Working Papers of the Institute of Management Berlin at the Berlin School of Economics and Law, paper no. 64.

[27] Höijer, B. (2010). "Emotional anchoring and objectification in the media reporting on climate change." Public Understanding of Science 19 (6): 717–731.

[28] Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K., & Johnson, C.A. (2001). "Climate Change 2001: The Scientific Basis." Cambridge University Press.

[29] IPCC (2013). "Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change." [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[30] Jung, J., Petkanic, P., Nan, D., Kim, J. H. (2020). "When a Girl Awakened the World: A User and Social Message Analysis of Greta Thunberg." Sustainability 12(7): 2707.

[31] Kaplan, A. M., Haenlein, M. (2010). "Users of the world, unite! The challenges and opportunities of social media." Business Horizons 53(1): 59–68.

[32] Kilian, T., Hennigs, N. (2011). "Unternehmerische Verantwortung zwischen Anspruch und Wirklichkeit: Eine empirische Analyse der Kommunikation CSR-relevanter Aspekte in Geschäftsberichten der DAX-30-Unternehmen von 1998–2009." Uwf UmweltWirtschaftsForum 19: 249–55.

[33] Kirilenko, A. P., Stepchenkova, S.O. (2014). "Public microblogging on climate change: One year of Twitter worldwide." Global Environmental Change 26(1): 171-182.

[34] Koenecke, A., Feliu-Fabà, J. (2019). "Learning twitter user sentiments on climate change with limited labeled data." arXiv preprint arXiv:1904.07342.

[35] Krabbe, O., Linthorst, G., Blok, K., Crijns-Graus, W., van Vuuren, D. P., Höhne, N., Faria, P., Aden, N., Carrillo Pineda, A. (2015) "Aligning corporate greenhouse-gas emissions targets with climate goals." Nature Climate Change 5: 1057– 60.

[36] Leas, E. C., Althouse, B. M., Dredze, M., Obradovich, N., Fowler, J. H., Noar, S. M., Allem, J.-P., Ayers, J. W. (2016). "Big Data Sensors of Organic Advocacy: The Case of Leonardo DiCaprio and Climate Change." PLoS ONE 11(8): e0159885.

[37] Levy, D. L. (2005). "Business and the evolution of the climate regime: The dynamics of corporate strategies." In D. L. Levy & P. J. Newell (Eds.), The Business of Global Environmental Governance.

[38] Lopes de Sousa Jabbour, A. B., Vazquez-Brust, D., Chiappetta Jabbour, C. J., Andriani Ribeiro, D. (2020). "The interplay between stakeholders, resources and capabilities in climate change strategy: converting barriers into cooperation." Business Strategy and the Environment 29(3): 1362–86.

[39] Mohammad, S., Turney, P. (2013). "Crowdsourcing a Word-Emotion Association Lexicon." Computational Intelligence 29(3): 436-465.

[40] Morales-Raya, M., Martín-Tapia, I., Ortiz-de-Mandojana, N. (2019). "To be or to seem: The role of environmental practices in corporate environmental reputation." Organization and Environment 32(3): 309–330.

[41] Müller, K., Pan, Y., Schwarz, C. (2023). "Social Media and Stock Market Participation." CEPR Discussion Paper No. 18445.

[42] Park, C. S., Liu, Q., Kaye, B. K. (2021). "Analysis of Ageism, Sexism, and Ableism in User Comments on YouTube Videos About Climate Activist Greta Thunberg." Social Media + Society, 7(3).

[43] Pearce, W., Holmberg, K., Hellsten, I., Nerlich, B. (2014). "Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report." PLoS ONE 9(4): e94785.

[44] Pham, H., Nguyen, V., Ramiah, V., Saleem, K., Moosa, N. (2019). "The effects of the Paris climate agreement on stock markets: evidence from the German stock market." Applied Economics 51(57): 6068–6075.

[45] Porter, M. E., Kramer, M. R. (2011). "Creating shared value: How to reinvent capitalism—And unleash a wave of innovation and growth." Harvard Business Review.

[46] Reid, E. M., Toffel, M. W. (2009). "Responding to public and private politics: Corporate disclosure of climate change strategies." Strategic Management Journal 30(11): 1157–78.

[47] Sabherwal, A., Ballew, M. T., van der Linden, S., Gustafson, A., Goldberg, M. H., Maibach, E. W., Kotcher, J. E., Swim, J. K., Rosenthal, S. A., Leiserowitz, A. (2021). "The Greta Thunberg Effect: Familiarity with Greta Thunberg predicts intentions to engage in climate activism in the United States." Journal of Applied Social Psychology 51: 321–333.

[48] Sarasini, S., Jacob, M. (2014). "Past, Present, or Future? Managers' Temporal Orientations and Corporate Climate Action in the Swedish Electricity Sector." Organization and Environment 27 3): 242–262.

[49] Sdrolia, E., Zarotiadis, G. (2019). "A Comprehensive Review for Green Product Term: From Definition To Evaluation." Journal of Economic Surveys 33(1): 150–178.

[50] Sharma, S., Nguan, O. (1999). "The biotechnology industry and strategies of biodiversity conservation: The influence of managerial interpretations and risk propensity." Business Strategy and the Environment 8(1): 46–61.

[51] Thunberg, Greta. (2019). "No one is too small to make a difference." Penguin.

[52] Tsalis, T. A., Nikolaou, I. E. (2017). "Assessing the Effects of Climate Change Regulations on the Business Community: A System Dynamic Approach." Business Strategy and the Environment 26(6): 826–43.

[53] UN Global Pulse (2015). "Using Twitter to Measure Global Engagement on Climate Change." Global Pulse Project Series, no.7.

[54] Weinhofer, G., Hoffmann, V. H. (2010). "Mitigating climate change - How do corporate strategies differ?" Business Strategy and the Environment 19(2): 77–89.

[55] Zhang, F., Zhu, L. (2019). "Enhancing corporate sustainable development: Stakeholder pressures, organizational learning, and green innovation." Business Strategy and the Environment 28(6): 1012–1026.

Appendix

Table A1: Classifier performance

Supervised classifier				
	Precision	Recall	F1	Support
0	0.88	0.91	0.89	348
1	0.92	0.88	0.90	373
Accuracy			0.90	721
Macro Avg.	0.90	0.90	0.90	721
Weighted Avg.	0.90	0.90	0.90	721
GPT 4o-mini classifier				
	Precision	Recall	F1	Support
0	0.89	0.97	0.93	1741
1	0.97	0.89	0.93	1862
Accuracy			0.93	3603
Macro Avg.	0.93	0.93	0.93	
Weighted Avg.	0.93	0.93	0.93	

Note: For the supervised classifier, metrics are computed using a held-out set of 20% of the manual labels. The GPT classifier metrics, on the other hand, are computed using the entire set of labels, as they are not otherwise employed in the classification process.

Table A2: Environment events by kind, count (2013-2023)

Event	# of occurrences
Flood (General)	532
Tropical cyclone	350
Riverine flood	242
Ground movement	143
Flash flood	140
Drought	117
Severe weather	67
Lightning/Thunderstorms	61
Blizzard/Winter storm	47
Tornado	38
Viral disease	32
Forest fire	29
Storm (General)	25
Bacterial disease	18
Wildfire (General)	17
Ash fall	16
Landslide (wet)	16
Cold wave	15
Severe winter conditions	14
Heat wave	10
Land fire (Brush, Bush, Pasture)	10
Hail	10
Extra-tropical storm	7
Lava flow	7
Mudslide	7
Volcanic activity (General)	6
Storm surge	4
Tsunami	3
Derecho	3
Pyroclastic flow	3
Glacial lake outburst flood	2
Collision	1
Avalanche (wet)	1
Coastal flood	1
Sand/Dust storm	1
	1995

Notes: Excludes occurrences in Germany. These refer to events in the January 1st 2013 - May 1st 2023 timeframe.

Table A3: Correlation matrices

Full timeline, weekly Green (2013-2023)

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	$\frac{Monday\ open_{i,t} - Friday\ close_{i,t}}{Friday\ close_{i,t}}$	1.0000									
(2)	$Greta\ treatment_t^{2+}$	-0.0299	1.0000								
(3)	$Greta\ treatment_t^{3+}$	-0.0454	0.9121	1.0000							
(4)	$Green_{i,t}^{week,\ 1+}$	-0.0103	0.1686	0.1539	1.0000						
(5)	$Green_{i,t}^{week,\ 2+}$	-0.0114	0.1461	0.1423	0.7169	1.0000					
(6)	$Green_{i,t}^{week,\ 3+}$	0.0045	0.1186	0.1181	0.5488	0.7655	1.0000				
(7)	$Green_{i,t}^{week,\ 4+}$	0.0042	0.0968	0.1003	0.4397	0.6134	0.8013	1.0000			
(8)	$Environment\ events_t$	-0.0537	0.2410	0.2424	0.1130	0.0865	0.0602	0.0405	1.0000		
(9)	$WE\ corp.\ env.\ tweets_{i,t}$	0.0024	0.0217	0.0258	0.1234	0.1459	0.1386	0.1487	0.0177	1.0000	
(10)	$Friday\ Close_{i,t}$	-0.0254	0.1836	0.1582	-0.0207	-0.0318	-0.0344	-0.0333	0.0833	-0.0228	1.0000

Greta timeline, weekly Green (2018-2023)

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	$\frac{Monday\ open_{i,t} - Friday\ close_{i,t}}{Friday\ close_{i,t}}$	1.0000									
(2)	$Greta\ treatment_t^{2+}$	0.0021	1.0000								
(3)	$Greta\ treatment_t^{3+}$	-0.0299	0.7984	1.0000							
(4)	$Green_{i,t}^{week,\ 1+}$	-0.0149	0.0006	0.0008	1.0000						
(5)	$Green_{i,t}^{week,\ 2+}$	-0.0191	0.0055	0.0223	0.7003	1.0000					
(6)	$Green_{i,t}^{week,\ 3+}$	-0.0002	0.0269	0.0405	0.5195	0.7419	1.0000				
(7)	$Green_{i,t}^{week,\ 4+}$	-0.0013	0.0419	0.0564	0.4061	0.5799	0.7817	1.0000			
(8)	$Environment\ events_t$	-0.0480	-0.0923	-0.0264	0.0432	0.0292	0.0129	0.0029	1.0000		
(9)	$WE\ corp.\ env.\ tweets_{i,t}$	-0.0034	-0.0002	0.0107	0.1017	0.1170	0.1032	0.1082	0.0079	1.0000	
(10)	$Friday\ Close_{i,t}$	-0.0111	-0.0148	-0.0274	-0.1606	-0.1443	-0.1141	-0.0981	-0.0085	-0.0410	1.0000

Full timeline, fortnightly Green (2013-2023)

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	$\frac{Monday\ open_{i,t} - Friday\ close_{i,t}}{Friday\ close_{i,t}}$	1.0000									
(2)	$Greta\ treatment_t^{2+}$	-0.0299	1.0000								
(3)	$Greta\ treatment_t^{3+}$	-0.0454	0.9121	1.0000							
(4)	$Green_{i,t}^{2week,\ 1+}$	-0.0199	0.1684	0.1550	1.0000						
(5)	$Green_{i,t}^{2week,\ 2+}$	-0.0139	0.1650	0.1564	0.7644	1.0000					
(6)	$Green_{i,t}^{2week,\ 3+}$	-0.0089	0.1603	0.1561	0.6342	0.8297	1.0000				
(7)	$Green_{i,t}^{2week,\ 4+}$	-0.0081	0.1495	0.1516	0.5442	0.7120	0.8581	1.0000			
(8)	$Environment\ events_t$	-0.0537	0.2410	0.2424	0.1272	0.1146	0.1032	0.0876	1.0000		
(9)	$WE\ corp.\ env.\ tweets_{i,t}$	0.0024	0.0217	0.0258	0.1050	0.1281	0.1291	0.1353	0.0177	1.0000	
(10)	$Friday\ Close_{i,t}$	-0.0254	0.1836	0.1582	-0.0165	-0.0320	-0.0419	-0.0410	0.0833	-0.0228	1.0000

Greta timeline, fortnightly Green (2018-2023)

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	$\frac{Monday\ open_{i,t} - Friday\ close_{i,t}}{Friday\ close_{i,t}}$	1.0000									
(2)	$Greta\ treatment_t^{2+}$	0.0021	1.0000								
(3)	$Greta\ treatment_t^{3+}$	-0.0299	0.7984	1.0000							
(4)	$Green_{i,t}^{2week,\ 1+}$	-0.0308	-0.0250	-0.0167	1.0000						
(5)	$Green_{i,t}^{2week,\ 2+}$	-0.0233	-0.0236	-0.0065	0.7526	1.0000					
(6)	$Green_{i,t}^{2week,\ 3+}$	-0.0179	-0.0028	0.0175	0.6154	0.8177	1.0000				
(7)	$Green_{i,t}^{2week,\ 4+}$	-0.0165	0.0065	0.0350	0.5216	0.6931	0.8475	1.0000			
(8)	$Environment\ events_t$	-0.0480	-0.0923	-0.0264	0.0530	0.0483	0.0455	0.0354	1.0000		
(9)	$WE\ corp.\ env.\ tweets_{i,t}$	-0.0034	-0.0002	0.0107	0.0857	0.1029	0.0978	0.1039	0.0079	1.0000	
(10)	$Friday\ Close_{i,t}$	-0.0111	-0.0148	-0.0274	-0.1722	-0.1797	-0.1674	-0.1509	-0.0085	-0.0410	1.0000

Table A4: Baseline - additional results

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.073	-0.050	-0.170***	-0.153***
	(0.046)	(0.042)	(0.038)	(0.035)
$Green_{i,t}$	-0.163	-0.161	-0.112*	-0.125
	(0.097)	(0.109)	(0.066)	(0.083)
$Greta\,treatment_t \times Green_{i,t}$	0.233**	0.209**	0.190***	0.184**
	(0.096)	(0.094)	(0.068)	(0.069)
$Env.\,events_t$	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\,corp\,env.\,tweets_{i,t}$	-0.012	-0.011	-0.011	-0.011
	(0.014)	(0.013)	(0.014)	(0.013)
$Friday\,close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	3+, 1 week	4+, 1 week	3+, 1 week	4+, 1 week
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.035	0.035	0.036	0.036
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least three or four tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table A5: Baseline, full timeframe - additional results

$Y = \frac{Monday\,Open_{i,t} - Friday\,Close_{i,t}}{Friday\,Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\,treatment_t$	-0.082***	-0.075***	-0.174***	-0.167***
	(0.027)	(0.026)	(0.025)	(0.025)
$Green_{i,t}$	0.009	0.016	0.011	0.012
	(0.046)	(0.050)	(0.038)	(0.044)
$Greta\,treatment_t \times Green_{i,t}$	0.041	0.011	0.045	0.025
	(0.050)	(0.056)	(0.042)	(0.048)
$Env.\,events_t$	-0.004***	-0.004***	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
$WE\,corp\,env.\,tweets_{i,t}$	0.001	0.001	0.001	0.001
	(0.014)	(0.014)	(0.014)	(0.014)
$Friday\,close_{i,t}$	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	3+, 1 week	4+, 1 week	3+, 1 week	4+, 1 week
$Greta\,treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2013-2023	2013-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	17,380	17,380	17,380	17,380
R^2	0.026	0.026	0.027	0.027
Total # firms	38	38	38	38

Notes: The sample consists of our full dataset (January 7th 2013 - April 24th 2023). $Greta\,treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least three or four tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table A6: Baseline, fortnights - additional results

$Y = \frac{Monday\ Open_{i,t} - Friday\ Close_{i,t}}{Friday\ Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\ treatment_t$	-0.149**	-0.118**	-0.218***	-0.211***
	(0.061)	(0.052)	(0.049)	(0.043)
$Green_{i,t}$	-0.282***	-0.275***	-0.194***	-0.226***
	(0.086)	(0.081)	(0.061)	(0.063)
$Greta\ treatment_t \times Green_{i,t}$	0.307***	0.287***	0.223***	0.254***
	(0.095)	(0.087)	(0.071)	(0.070)
$Env.\ events_t$	-0.004**	-0.004**	-0.003*	-0.003*
	(0.002)	(0.002)	(0.002)	(0.002)
$WE\ corp\ env.\ tweets_{i,t}$	-0.011	-0.011	-0.010	-0.010
	(0.013)	(0.013)	(0.013)	(0.013)
$Friday\ close_{i,t}$	-0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	3+, 2 weeks	4+, 2 weeks	3+, 2 weeks	4+, 2 weeks
$Greta\ treatment_t$	2+, weekend	2+, weekend	3+, weekend	3+, weekend
Sample	2018-2023	2018-2023	2018-2023	2018-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	8,514	8,514	8,514	8,514
R^2	0.036	0.036	0.037	0.037
Total # firms	38	38	38	38

Notes: The sample consists of the timeframe in which Greta Thunberg is active on Twitter (June 18th 2018 - April 24th 2023). $Greta\ treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least three or four tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

Table A7: Baseline, fortnights (full timeframe) - additional results

$Y = \frac{Monday\ Open_{i,t} - Friday\ Close_{i,t}}{Friday\ Close_{i,t}}$	(1)	(2)	(3)	(4)
$Greta\ treatment_t$	-0.082***	-0.075***	-0.174***	-0.167***
	(0.027)	(0.026)	(0.025)	(0.025)
$Green_{i,t}$	0.009	0.016	0.011	0.012
	(0.046)	(0.050)	(0.038)	(0.044)
$Greta\ treatment_t \times Green_{i,t}$	0.041	0.011	0.045	0.025
	(0.050)	(0.056)	(0.042)	(0.048)
$Env.\ events_t$	-0.004***	-0.004***	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
$WE\ corp\ env.\ tweets_{i,t}$	0.001	0.001	0.001	0.001
	(0.014)	(0.014)	(0.014)	(0.014)
$Friday\ close_{i,t}$	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
$Green_{i,t}$	3+, 2 weeks	4+, 2 weeks	3+, 2 weeks	4+, 2 weeks
Sample	2013-2023	2013-2023	2013-2023	2013-2023
Fixed effects	Firm, month, year	Firm, month, year	Firm, month, year	Firm, month, year
Total # observations	17,380	17,380	17,380	17,380
R^2	0.026	0.026	0.027	0.027
Total # firms	38	38	38	38

Notes: The sample consists of our full dataset (January 7th 2013 - April 24th 2023). $Greta\ treatment_t$ is equal to 1 when Greta Thunberg tweets at least twice, in specifications (1) and (2), or three times, for (3) and (4), about the environment in the respective weekend and 0 otherwise. $Green_{i,t}$ is equal to 1 when a firm has made at least three or four tweets about the environment in the 7 days leading to the respective Friday and 0 otherwise. Firm, month and year fixed effects are employed and standard errors are clustered at the firm level throughout. ***, **, and * denote significance at the 1%, 5% and 10% level, respectively. Coefficients and standard errors are multiplied by 100.

CHAPTER THREE

Audiovisual speaker recognition, diarization and transcription: towards an autonomous application to election debates*

Henrique Alpalhão

Abstract

Recent times have seen substantial and accelerating technological advances in topics such as large language models (LLMs), machine and deep learning. Many of these developments have the potential to revolutionize data collection methods and open up to detailed, objective analysis material that was previously either not interpretable in such a way or strenuously hard to. In this work, I propose to leverage one such development – deep-learning and LLM-based audio and video analysis – to facilitate the analysis of a source of plentiful data: election debates and politician interviews. I describe a method that can be applied to material in any language, requires relatively light prep-work and delivers full, highly accurate diarized video transcriptions. I further propose an expansion that firstly allows the pipeline to autonomously learn how speakers sound from only a small set of pictures and the material to be analyzed itself; and secondly is able to learn from experience to become more accurate the more material it encounters. These diarized transcriptions can be explored for several quantitative and qualitative questions and are easily interpreted by LLMs to obtain measures as diverse as political actor demeanor, policy positions, party or left-right alignment, and the tracking of all these across time and countries; and these analyses will be replicable in ways that previous, expert-led methods could not be. A brief application, focusing on identifying the policy foci of different Portuguese politicians from their election debate participations, illustrates the method.

1 Introduction

It is difficult to find a field that stands to benefit as much as political science from improvements in data collection and access to new or more granular data sources. Indeed, even quantitative work often requires grounding on points of departure that are qualitative, potentially subjective or reliant on an accumulated memory and experience with the topics at hand. This is the case, for example, for the numerous enterprises that require manual classifications (e.g. party classification over ideological spectra) or expert surveys (e.g. data derived from political manifesto analysis). The fact that experts and authors might not always agree, and that the opinions that substantiate these decisions are subject to evolve across time, can create challenges of consistency across time and the corpus of literature, slow the achievement of consensus where it is feasible and desirable, and hamper replicability. These issues, of course, also create repercussions wherever political science might be put to practice (political communication services, politicians themselves, voters, journalists), making their addressing an especially relevant enterprise.

*This version: November 2025.

The "fundamental scientific objective of specifying reliable and replicable methods for collecting the data needed" (Benoit et al., 2016) is thus an open pursuit in the political science literature.[1] In this work, I propose to leverage recent developments in the field of computerized speech analysis to tackle this issue, providing a building block for more objective and reproducible data-collection regarding political phenomena. I design a simple and easily replicable pipeline that takes TV-studio environment politician videos (e.g. interviews, debates) and an index of speakers. With these two inputs, it automatically and reliably identifies politicians by both voice and face, transcribing and assigning them their spoken lines. In the age of large language models (LLMs), these transcripts may be passed to one such model to obtain replicable analysis that can take into account more data than an expert is likely to keep in mind at each time.

In its basic state, my method relies mostly on audio diarization, resorting to video only as a fallback when the audio track cannot confidently be assigned to a known speaker. I proceed to provide extensions that introduce several improvements: firstly, an automated collection of audio samples through face recognition facilitates index-building, significantly reducing the amount of required manual steps. Secondly, a self-learning logic that works such that whenever a specific speaker is identified, any deviations from their sound template can be incorporated into the index, making the process more precise the more it encounters the population under study. Thirdly, an LLM-powered correction process that uses context to infer previously-unidentified speaker identities to a very competitive accuracy level.

This work will describe this proposed method in an iterative way. I begin with a literature review, in section 2, that describes the application of diarization methods in the literature and how analogous data is treated in the political science corpus. Section 3 describes, in general, how the method works and what it requires. Section 4 describes the "skeleton" pipeline: an audio diarization with video fallback. Section 5 progresses by introducing the self-learning modules of automated voice bank creation and enrichment, as well as the LLM-based correction method. Section 6 presents accuracy metrics for these methods, while section 7 illustrates their usage through an analysis of policy foci obtained from the diarized transcripts of a debate corpus. Finally, section 8 concludes with a brief recap and a few suggestions for further research.

2 Theory and literature review

Text is an important type of data in political science: it is a direct input for pursuits such as political manifesto, written media and social media analysis, and an intermediate product for all others that must be translated into text before they can be processed (e.g. broadcast or parliamentary debates, media and social media in video or audio format). While this work focuses on the more specific pursuit of generating data from spoken sources, the end result – a diarized transcript – is still text data. To understand its idiosyncrasies and how to employ it, it is thus useful to look at the literature that works with text. This section does that first, before moving on to a review of the speaker diarization literature.

2.1 Text-as-data in political science

In the last decade, pursuits that rely on text have experienced a sizable boom, both due to increased data availability and improved methods for its analysis (Wilkerson and Casas, 2017). In their 2003 paper, Laver et al. describe their analysis as one that "treats text not as discourses to be understood and interpreted but rather, as data in the form of words", establishing a dichotomy between a qualitative, document-level analysis of meaning and a quantitative, word-level "scoring" approach. With the advent of LLMs and similar technology, this distinction becomes significantly blurred as a computerized, relatively replicable interpretative method can be applied to text

[1] See also Zulianello (2014) for specific concerns with the commonly-used Comparative Manifesto Project (https://manifesto-project.wzb.eu/) data.

at scale. This is a nascent application, but some examples, such as Kato et al. (2024), Ceron et al. (2024), and Di Leo et al. (2025) can already be found.

Several media have acted as the source of text data for political scientists. One popular such origin, for example, is Twitter (now X)[2] data, be it on the voter or politician side. Barberá and Rivero (2015) provide a good literature review of work focusing on this source. Note, however, that a more restrictive academic access policy in the X era has hampered the availability of these data, and thus academic production in the field. Ernst et al. (2017), Theocharis et al. (2020), and Silva and Proksch (2022) are examples of work that use tweets as their main data source.

Another strand of the literature delves into parliamentary (i.e. legislative) debates, which Fernandes et al. (2021) review. They discuss how these provide useful gauges for party positions in the periods between elections and, in general, provide opportunity for politicians to show responsiveness to their electorate; they are also likely to carry information that might not be revealed in static sources like party manifestos, as those are not subject to the cross-examination and dialogue dynamics that political debate provides. Chaqués-Bonfort and Baumgartner (2013) further find that those in opposition are more likely to use them to drum up media interest in a specific topic, while Bevan and John (2016) argue that those in government use them to highlight the executive's achievements.

In the specific case of this article, I turn to electoral debates – meaning broadcast debates between two or more candidates for a given election – rather than their parliamentary analogue for data. Public and academic interest in electoral debates date at least to the first Kennedy vs. Nixon presidential debate in 1960 – the first of its kind in a US election – which is often cited as a pivotal moment and mentioned in studies to this day (e.g. Druckman, 2003; Self, 2005; Bidwell et al., 2020; Herbeck and Drudy, 2022). This interest has been persistent across time: McKinney and Carlin (2004) state that US presidential debates were, at the time, the political event that reached the most people, and Bidwell et al. (2016) discuss how the first Clinton vs. Trump debate, in 2016, had been the most-viewed American debate so far (with 84 million viewers).[3] Bidwell et al. (2020) further discuss how they allow for a relative evaluation of policy proposals, competence and charisma between participating candidates and provide equal opportunities for challengers and incumbents; also noting how eagerly debates are discussed and dissected by the media, often with hours of media commentary and candidate performance analysis attached to each.

Given this, it is no surprise that some literature focuses on electoral debates. Clementson and Eveland (2016), for example, examine US presidential debate and press conference Q&As to evaluate question-dodging behavior, departing from official transcripts (which are often not broadly available). They find that, in debates, politicians are more likely to dodge questions by introducing new topics rather than outright refusing to answer. Bidwell et al. (2020), on the other hand, use public debate screenings in Sierra Leone to experimentally test their impact on voters and politicians. They find that being exposed to debates improves voters' political knowledge, impacting their vote, candidate behavior during campaigns, and pressure for the accountability of elected official spending.

2.2 Speaker diarization and recognition

"To diarize" is a less common verb that means "to record in a diary". Speaker diarization can be defined as the attempt to answer the "who spoke when" (Anguera et al., 2012) and "what was spoken" questions in regard to a given audio or audiovisual material (Kynych et al., 2024). Tranter and Reynolds (2006), Anguera et al. (2012) and Park et al. (2022) are good sources for a historical overview of this method's development and a review of the current state of the art. Diarization technology dates back to the 1990s, when it focused on automatic speech recognition for air traffic control (Gish et al., 1991; Siu et al., 1992) and broadcast news (Ajmera and Wooters, 2003; Meignier et al., 2006). Other applications have included meeting (van Leeuwen and Konečný, 2007; Vijayasenan

[2]https://x.com.

[3]Which, according to The Guardian, dropped to 67 million for the Harris-Trump 2024 showdown (https://www.theguardian.com/us-news/2024/sep/12/us-presidential-debate-tv-ratings-harris-trump-abc). For more figures, see https://www.statista.com/chart/23075/estimated-tv-viewership-of-presidential-debates/.

et al., 2009) and telephone (Zhu et al., 2005; Meignier et al., 2006) conversations. Park et al. (2022) discuss how, more recently, deep learning has become an integral part of the field's evolution, allowing for performance gains, increased robustness against diverse acoustic conditions, and facilitated large-N training.

As a manual exercise, diarization is a quite expensive and time-consuming ordeal – as such, the automation of this process has received significant and increasing attention; driven in part by the current-day proliferation of new audiovisual content, and in part due to an increased ability for and inclination towards digitizing TV archives (Mingote et al., 2024). When using video material as inputs, it can be approached in two main ways: either by focusing on audio only or on both audio and image.

Using only audio is a viable option for several reasons. Firstly, it is less computationally intensive, making outputs quicker and cheaper to obtain (Mingote et al., 2024). Secondly, it has the most accumulated experience: the first audio diarization efforts date back to the 1990s. Thirdly, it is more likely to work as a stand-alone method – while video images by themselves will never capture all the necessary information to infer who an off-camera speaker is (Kynych et al., 2024), or to distinguish an off-camera speaker from silence, an accurate enough audio approach is able to do so by identifying a change in speaker voice characteristics.

Nonetheless, Park et al. (2022) point joint audiovisual modeling as an important avenue for improvement of these methods. A video's image does contain useful data that can be used to complement an audio approach, such as face and lip movement, spatial positioning, and gestures (Kynych et al., 2024, Mingote et al., 2024).[4] While earlier literature focused mainly on lips, more recent efforts, such as Sharma and Narayanan (2022), incorporate broader cues. Recording conditions further play an important role in deciding the kind of diarization to implement: pure audio-based diarizarion is less effective the more diverse acoustic conditions are and the more audio confounders exist. Background music, overlapping speech, sound effects and intra-speaker variability, as well as frequent speaker change, all hamper the accuracy of pure-audio methods (Sharma and Narayanan, 2022, Mingote et al., 2024). Mingote et al. (2024) is another example of a paper that proposes an integrated audio-visual diarization framework, providing an interesting literature review of previous applications of this method.

Finally, the prior knowledge of speakers is not required to perform diarization (Park et al., 2022). However, such knowledge can be leveraged to individually identify speakers, rather than just separate them, by matching them to a databank. Mingote et al. (2024) note how this approach is not often discussed in the literature and implement it, a course I also follow in this work. Still on this topic, Kynych et al. (2024) find that the combination of audio and image does help disentangle overlapping speech.

Software optimization has also been an important concern in the literature. While not of the utmost relevance for offline applications (as is my case), as input processing can take place before analysis, the several real-time applications for these methods (e.g. online subtitling) require pipelines that work as efficiently as possible (Kynych et al., 2024). Hardware considerations are also closely connected to this discussion as the other side of shortening pipeline runtimes. In general, they can be run either on CPUs or GPUs, with the former being cheaper and more readily available and the latter providing significant speed improvements. For real-time outputs, Kynych et al. (2024) recommend CPU operation due to "operational and deployment costs". They propose an audio-only and an audiovisual real-time diarization method, with the former running on CPU and the latter requiring GPU acceleration.

2.3 From parliamentary to electoral debates

While the similarities between parliamentary and electoral can be numerous – namely regarding the protagonists, the kind of language used, and at least occasionally the topics under discussion – their differences are also worthy of note and mean that their treatment and analysis should be treated as different, if comparable, efforts. Firstly, parliament sees many more individuals speaking, both in general and per session – while, for election debates, one

[4]For more on the association between vocalizing and facial behavior, refer to Yehia et al. (1998).

can keep a database of only party leaders and journalists, for their parliamentary counterpart they will need to account for hundreds of potential speakers.

Secondly, discussion in parliament is made between politicians, who are generally particularly skilled speakers stemming from relatively homogeneous environments; while in debates politicians speak as much to their opponent as to the audience. This will mean that both the form of speech and its content will be different across the two formats. Thirdly (and closely related to the last point), parliamentary and electoral debates have different and possibly disconnected objectives: the former is meant to design, perfect and implement policy, while the latter is meant to convince voters. In the era of immediate politics through social media, however, these differences are likely to progressively blur as parties begin to treat parliament as an additional way to feed their content machine and generate clips with which to directly reach their voters.

It is also worth noting that parliamentary debates are structured affairs, implementing allotted speaking times, required yielding and the like. Electoral debates see moderators attempt to establish a similar playing field, but nonetheless often degenerate into a more chaotic brand of rhetoric in which certain kinds of politicians might thrive – this means an average parliamentarian might make a mighty debate opponent; and that transcribing these items will pose different challenges. Finally, parliaments also generally make a broad range of resources available, from full professional transcripts to extensive MP and institutional role databases, that must be obtained piecemeal or tailor-made for electoral debate applications.

In practice, these differences suggest that electoral debates not only require original supporting datasets, but also pose particular challenges. More frequent speaker overlap; the occurrence of audible interruptions; shorter average interventions; and a higher range of displayed emotions are all differentiating characteristics of these debates, and a competent transcription and diarization strategy will have to address them.

3 Building blocks

Having discussed the motivation and literature, I move to the substance of this article – designing a diarization method and implementing it through an example set of inputs.

3.1 Obtaining the inputs

Naturally, one should begin by identifying and collecting the files they intend to analyze. The pipelines I am proposing are tailored for studio conversation environments, and as such excel at analyzing debates and interviews. With some changes, this method could also be applied to broader set of circumstances, but adjustments would have to be made to account for more complex acoustic conditions (e.g. background noise, the lack of individual microphones). For the purpose of this work, I collected Portuguese election debates for six elections: the legislative elections of 2019, 2022 and 2024, the European elections of 2019 and 2024, and the presidential elections of 2021.

Generally, Portuguese TV electoral debates follow a consistent model. For each election, all party leaders of running parties who are currently represented in Parliament participate in extensive one-on-ones with all other party leaders in the same circumstances. These debates last for around half-an-hour and are moderated by a single journalist, except for the PS-PSD debate in legislative elections, which usually lasts from 60 to 90 minutes and is moderated by a group of three journalists.[5] In addition to these, three multi-party, 2-hour debates are also usually held: two between all the previously-mentioned parties, one of which is broadcast on TV and the other on the radio, and a third one between all running party leaders who are not currently represented in Parliament. An exception to

[5]This is due to these two parties, traditionally, being the most voted. While PS was still the second-most voted party in the 2025 legislative elections, their votes only translated into the third-largest parliamentary representation (58 MPs vs. Chega's 60), which could lead to changes to this model in the future.

Table 1: Debate index example

Date	P1	P2	S1	S2	Moderator	Election	Link	ID
05/02/2024	PS	IL	Pedro Nuno Santos	Rui Rocha	Clara de Sousa	Leg24	https:/...	Leg24_1
05/02/2024	CH	PAN	André Ventura	Inês Sousa Real	João Adelino Faria	Leg24	https:/...	Leg24_2
06/02/2024	AD	BE	Luís Montenegro	Mariana Mortágua	Sara Pinto	Leg24	https:/...	Leg24_3
06/02/2024	CH	IL	André Ventura	Rui Rocha	Rosa de Oliveira Pinto	Leg24	https:/...	Leg24_4
06/02/2024	CDU	PAN	Paulo Raimundo	Inês Sousa Real	João Adelino Faria	Leg24	https:/...	Leg24_5
07/02/2024	IL	L	Rui Rocha	Rui Tavares	João Póvoa Marinheiro	Leg24	https:/...	Leg24_6

Notes: The "Link" column is abridged for space savings. All of the entries above can be found at https://www.rtp.pt/play/p12901/debates-legislativas-2024-tvicnn, https://www.rtp.pt/play/p12899/debates-legislativas-2024-sicsic-noticias, or https://www.rtp.pt/play/p12900/debates-legislativas-2024.

this rule occurred in the 2024 European elections, for which parties could only agree to four-way debates between the parties with previous representation.

Given the particular challenges that debates with many speakers present (e.g. much more frequent speaker changes, interruptions by several different voices, lower relative speaking time per speaker), I exclude the large, multi-party debates from my sample, keeping, however, the 2024 European debates as a stress test. This yields a dataset with 108 videos and a total runtime of 62 hours, 40 minutes and eight seconds (at an average 34 minutes and 49 seconds per video). The excluded set consists of 27 videos (note most of the large debates are broadcast in separate parts – this equates to 15 debates) and a total runtime of 29 hours, 33 minutes and 16 seconds.

The first step is to make a list of these debates and find a source for them – in my case, the Portuguese public broadcaster (RTP – Rádio e Televisão de Portugal) makes them freely available. Table 1 illustrates an index that serves this purpose.

With this index in hand, I use a bulk collection auxiliary Python pipeline, relying on the yt_dlp package,[6] as presented in appendix A1. This step creates a local copy of all the videos of interest and of their separate audio tracks.

3.2 Building a people index

The second step is to build a "people index": a list of the individuals I want the pipeline to recognize. In my case, this comprises everyone who appears in the videos I collected – the politicians who participated in the debates and the journalists who acted as moderators. For each of them, it is then necessary to collect both images and voice samples.

For images, I manually obtained samples from the highest-resolution videos in my database, taking care to include at least a frontal shot and a shot of each side. It possible to collect as many additional samples as desired, and facial detection accuracy should improve with the number of images. For this exercise, I collected either four or five images per person.[7] For voice samples, on the other hand, I collected roughly a minute of speech per individual, broken up into 10-second segments. To this end, I depart from the audio tracks obtained from the pipeline in A1, manually find uninterrupted 10s segments for each speaker, and export them in WAV format. This collection is done using Audacity[8] and is the most time-consuming part of the required preparation – as such, section 6 discusses a method that automates this process. Both the image and audio files collected in this section are saved in the format "NameSurname#" (e.g. MarisaMatias1.png; MarisaMatias2.wav) in their own folder in the working directory. Table

[6] https://github.com/yt-dlp/yt-dlp.
[7] Video stills are obtained directly from Quicktime and exported to PNG format through Pixelmator.
[8] https://www.audacityteam.org/.

Table 2: People index example

Name	Party	Front	Side1	Side2	...	Voice1	...
Marisa Matias	BE	MarisaMatias1	MarisaMatias2	MarisaMatias3		MarisaMatias1	
Mariana Mortágua	BE	Mortagua1	Mortagua2	Mortagua3		Mortagua1	
Catarina Martins	BE	CatarinaMartins1	CatarinaMartins2	CatarinaMartins3		CatarinaMartins1	
Francisco Rodrigues dos Santos	CDS	Chicao1	Chicao2	Chicao3		Chicao1	
Assunção Cristas	CDS	Cristas1	Cristas2	Cristas3		Cristas1	
Nuno Melo	CDS	Melo1	Melo2	Melo3		Melo1	

Notes: Some image and audio columns are omitted for space savings. They are as follows: Extra1; Extra2; Voice2; Voice3; Voice4; Voice5; Voice6.

2 exemplifies this index.

4 Audio diarization with video fallback

With manual collections sorted, one can move to their processing. I begin by describing the baseline method: an audio diarization pipeline with video fallback. This requires the elements described in the previous section: the input videos, the people index, and the picture and voice samples. Figure 1 describes the process visually.

4.1 Video input to transcript: Whisper

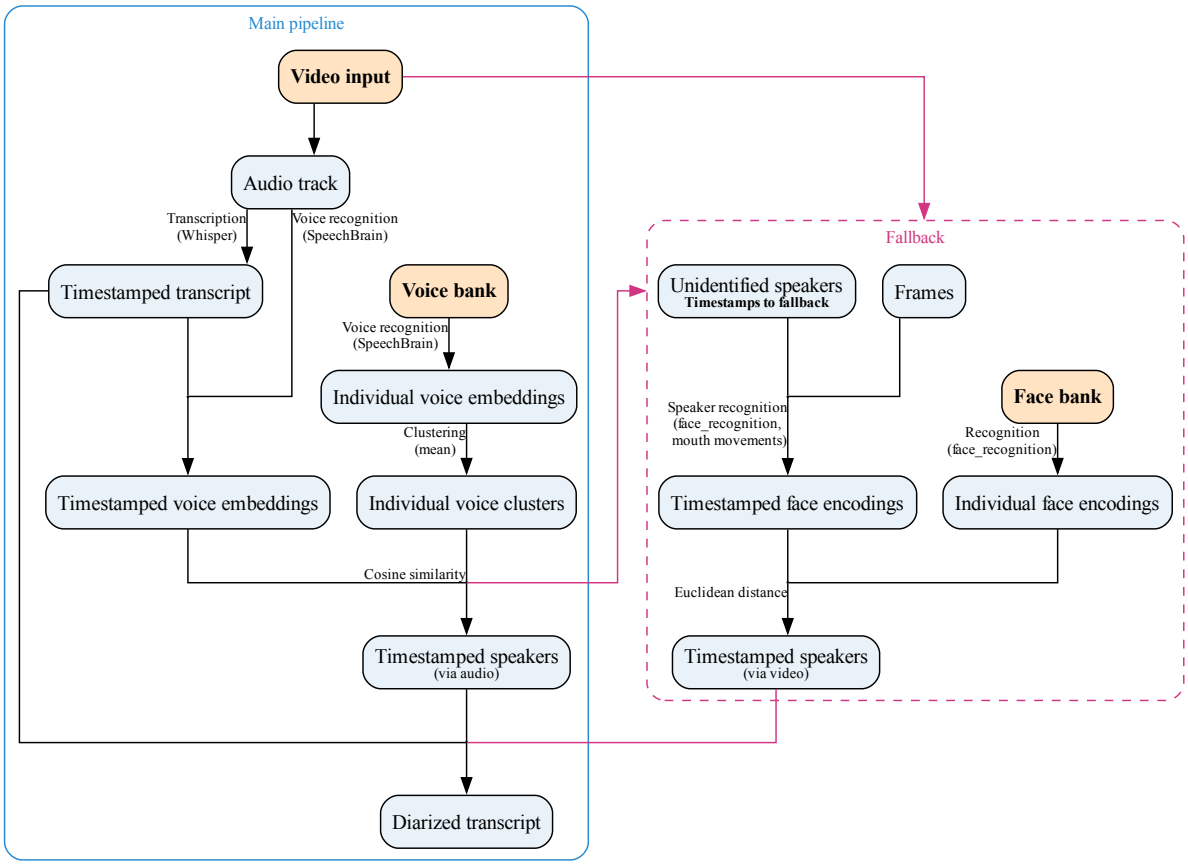
The process begins by taking the supplied video file and extracting its audio track via Pydub and FFmpeg.[9] This audio track is then passed on to Whisper, a deep learning OpenAI speech recognition package,[10] which produces a segmented, timestamped transcription that follows the rhythm of speech: lines which are spoken fluidly and continuously tend to be grouped into a single segment, while those interrupted by pauses or speech crutches are more likely to be split up into several segments. In principle, each segment is spoken by a single individual, but the same speaker can utter several consecutive segments. Settings for this step are as follows:

```
def whisper_segments(wav, model="large", lang="pt"):
    model=whisper.load_model(model, device=DEVICE_WHISPER)
    res = model.transcribe(
    wav, task="transcribe", language=lang,
    temperature=(0.,), beam_size=5,
    no_speech_threshold=0.8, logprob_threshold=-0.5,
    compression_ratio_threshold=2.8,
    condition_on_previous_text=False)
    return [{"start":s["start"],"end":s["end"],"text":s["text"]}
            for s in res["segments"]]
```

Several relevant design choices can be discussed here. Firstly, I opt for the large Whisper model – its most accurate, but also slower implementation. For applications that require quicker transcription, the "turbo" model is likely a good alternative.[11] I also force Whisper to recognize the language as Portuguese through *lang="pt"*. This is a safe choice if we know dialogue to be exclusively in a given language, but can be relaxed for multilingual or unknown-language environments. Through *task="transcribe"*, I instruct Whisper to output text in the same language as it is spoken; its alternative *"translate"* will output English translations instead. This option might be

[9] https://pydub.com/, https://ffmpeg.org/.
[10] https://github.com/openai/whisper.
[11] See https://github.com/openai/whisper/blob/main/model-card.md for all alternatives.

Figure 1: Baseline pipeline



Notes: Diagram representation of the baseline pipeline and its fallback process. The fallback branch only runs when a segment cannot be confidently attributed to a known speaker via the main branch, and only for those segments.

well applied, for example, in a bulk translation of material from different countries for comparative analysis. As timestamps remain unaffected by this choice, the rest of the pipeline may be used unaltered to this end.

By setting *temperature=(0.,)*, I obtain a deterministic output, meaning that multiple runs will always generate the same transcript for the same input. This can be set to higher values (to a maximum of 1), which will yield higher-randomness, less coherent outputs that are often described as "more creative" (Peeperkorn et al., 2024). While those are generally not desirable characteristics for this specific pipeline, if computing time is not an issue one can leverage this function by passing N temperature values (e.g. *temperature=(0.0, 0.2, 0.3)*). In this case, whenever Whisper produces an interpretation of a 30-second window that it is not confident in, it drops that transcription and tries again with the next temperature value (in case no transcription meets the criteria, it will return the best one); this is likely to improve the accuracy of the final transcription. With *beam_size=5*, on the other hand, Whisper will consider five different possible transcriptions for each window, selecting the most likely one. With *temperature=(0.,)*, this becomes especially relevant as it is the only source of alternative outputs (which remain deterministic and reproducible). Note that, when several temperature values are supplied, *beam_size* applies to each – *temperature=(0.0, 0.2, 0.3)* and *beam_size=5*, for example, will yield up to 15 different hypotheses for each window, significantly increasing runtime.

The three following knobs – *no_speech_threshold=0.8*, *logprob_threshold=-0.5*, *compression_ratio_threshold=2.8* – control how Whisper evaluates each transcription's likelihood. *no_speech_threshold* is a voice activity detector, controlling how likely it is to dismiss a window as only silence. The default value is 0.6, by raising it to 0.8 I account for the fact that complete silence in such a long window is unlikely for my setting. The remaining two only apply to the multiple temperatures case, dictating whether the current value is kept or the next value is tried: l*ogprob_threshold=-0.5* dictates that the average log-probability of all tokens in the window under analysis must be at least $e^{-0.5} \approx 61\%$; *compression_ratio_threshold=2.8* is a parameter that measures the ratio of "useful" to "useless" text (e.g. words vs. utterances such as "uh" or "eh" or textual/punctuation hallucinations).

Finally, *condition_on_previous_text=False* tells Whisper not to use the previous window to inform its transcription of the current window. Setting this to *true* allows the model to infer what is being said from context, which can be useful for lower-quality audio and shorter sentences but is notoriously poor at transcribing the longer-form speech that can usually be found in a debate. It also noticeably increases the amount of hallucinations and occasionally even sends Whisper into a monologue-like loop where it produces an initial hallucination and replies to itself several times. For these reasons this knob is best set to *False* for this application.

4.2 Voice embeddings: SpeechBrain/ECAPA and sample matching

With the timestamped transcript in hand, we turn our attention to voice recognition. This process begins with the reduction of the voices heard in each timestamp to a vector representation, for which I employ SpeechBrain.[12]

For this application, I am using a ECAPA-TDNN model (Emphasized Channel Attention, Propagation and Aggregation – Time Delay Neural Network).[13] This specific model is pre-trained on the VoxCeleb1 and VoxCeleb2 datasets, which are drawn from YouTube content and, combined, include more than a million utterances by more than 7000 celebrities covering multiple different languages and accents (Nagrani et al., 2017; Chung et al., 2018). In principle, it would be possible to train a new model using original samples (and a significant amount of time), but given VoxCeleb's broad scope it would be unlikely to yield more accurate results. Briefly, this module receives an audio snippet as input, detects voice characteristics within and reduces them to a 192 dimension vector – the "embedding". The functions that slice the audio input to obtain the relevant snippets and compute embeddings – seg_emb and compute_spk_emb – can be found in appendix A2.

[12] https://speechbrain.github.io/. In my application, it is called as spkrec=SpeakerRecognition.from_hparams("speechbrain/spkrec-ecapa-voxceleb", run_opts={"device":DEVICE_ECAPA}).

[13] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb. For details on the workings of this kind of model, refer to Desplanques et al. (2020) or Dawalatabad et al. (2021).

The same process is applied to the pre-collected voice samples: all samples are collapsed to embeddings and a centroid is computed per individual, creating a representation of how that speaker sounds. The function that performs this task can be found in appendix A3. With these in hand, each segment's ECAPA representation is compared to the people centroids, and is matched to the most similar one, as long as it displays a minimum degree of similarity – in my application a threshold of at least 40% cosine similarity between each segment's vector and the speaker representation is applied. Those that cannot be confidently matched to anyone are labelled as "Unknown", and all others are assigned a speaker. This function is presented in appendix A4.

Finally, a subprocess identifies the prevalence of each speaker. Those who speak for less than 10% of the video's running time are labelled as "rare" and merged into the closest "major" speaker as evaluated by their ECAPA embeddings. This step ensures that occasional aberrant segments do not mislead the audio detection – for example, a given speaker's cough might disturb their embedding enough that the closest centroid becomes someone else, who is not present in the video under analysis. Given these phenomena only happen occasionally, this step sweeps them up and reassigns them correctly. It is worthy of note, though, that the rarity threshold (10% in this example) might need to be tailored to each application. 10% works well for three-person debates, but tends to eliminate correct identifications when more speakers are present (e.g. 10 speaker-debates), and thus should be lowered. This final helper function can be found in appendix A5.

4.3 Video fallback

At this stage, the pipeline progresses to the video fallback. It begins by identifying all segments with a yet-unidentified speaker – meaning those for which audio recognition was unsuccessful – and extracting only the frames corresponding to those timestamps through OpenCV.[14] Applying this step only to the relevant frames is an important time-saving step, as facial recognition is a computationally-intensive process.

I then use the face_recognition package[15] to identify faces in those frames. For each detected face, it returns its location within the frame, a 128-dimension face encoding and a dictionary of 68 facial landmarks, which can be used to locate facial features (e.g. the mouth, eyes, etc.). At this stage, encodings are also obtained for the face samples in the people index, creating a vector representation of each known individual's likeness.

Once this is done, the detected faces are matched to the indexed ones. If several known faces are sufficiently similar,[16] the most similar one is matched and identified as the person on screen. If none meet the similarity threshold, the "Unknown" label is retained. Two guardrails are in place for this process: firstly, faces in the video whose height or width are smaller than 10% of the frame's corresponding measures are ignored – this is useful to exclude sign language interpreters and wide shots, which might confuse the detection due to the presence of many potential speakers and the lower amount of information than further-away faces display. Secondly, only speakers who were already identified in the audio step are whitelisted for matching, which serves as an automatic guide for the pipeline that has significant impact on detection accuracy.

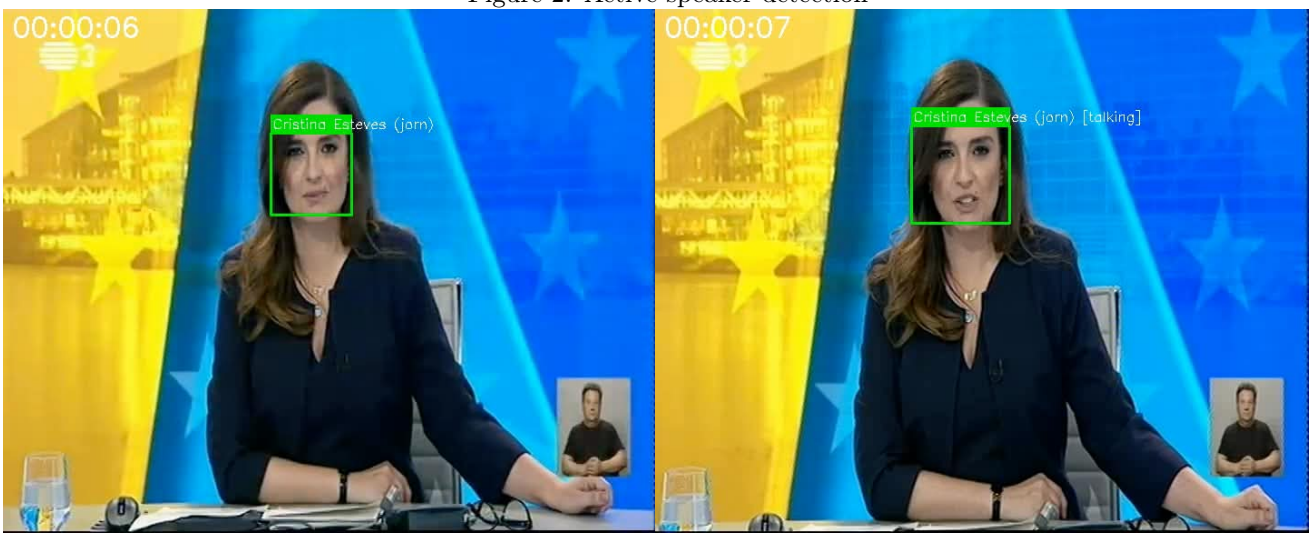
Finally, the function uses the facial landmarks pertaining to the mouth to identify, for each analyzed frame, whether the detected individual is speaking or not. It detects a given face as having their mouth open if lip separation equals at least 10% of the mouth's total width. Using this, it credits a detected face as speaking only after three consecutive frames with an open mouth, and infers the end of speech after three consecutive mouth-closed detections. At the end of this process, the spoken segment under analysis is attributed to the person on-screen who was detected as speaking the longest within the relevant timestamps, retaining its "Unknown" label if no active speaker can be identified. Figure 2 illustrates these two states with two debate frames annotated directly by the

[14] https://opencv.org/.

[15] https://github.com/ageitgey/face_recognition. For more in depth information, see the documentation at https://app.readthedocs.org/projects/face-recognition/downloads/pdf/stable/ and Adam Geitgey's article at https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78.

[16] For my application, I am using a maximum Euclidian distance of 0.5, which is slightly more severe than the commonly-used 0.6.

Figure 2: Active speaker detection



Notes: Frames obtained from the May 2nd 2019 debate between Nuno Melo (CDS-PP) and Marisa Matias (BE), moderated by Cristina Esteves, during the campaign for the 2019 European elections.

pipeline.

Once all of this is done, the process outputs four files: the full diarized transcripts in JSON and .txt, a list of identified speakers in .txt format that is useful to quality assure speaker recognition, and an .srt file that can be used as subtitles in conjunction with each input video to quality assure transcriptions. The code for this sub-pipeline can be found in appendix A6. For this baseline, 2.72% of segments are classified as having an unknown speaker (0.99% of spoken time and 1.07% of spoken words), and video identification is used for 3.89% of all segments (1.44% of spoken time and 1.58% of spoken words).

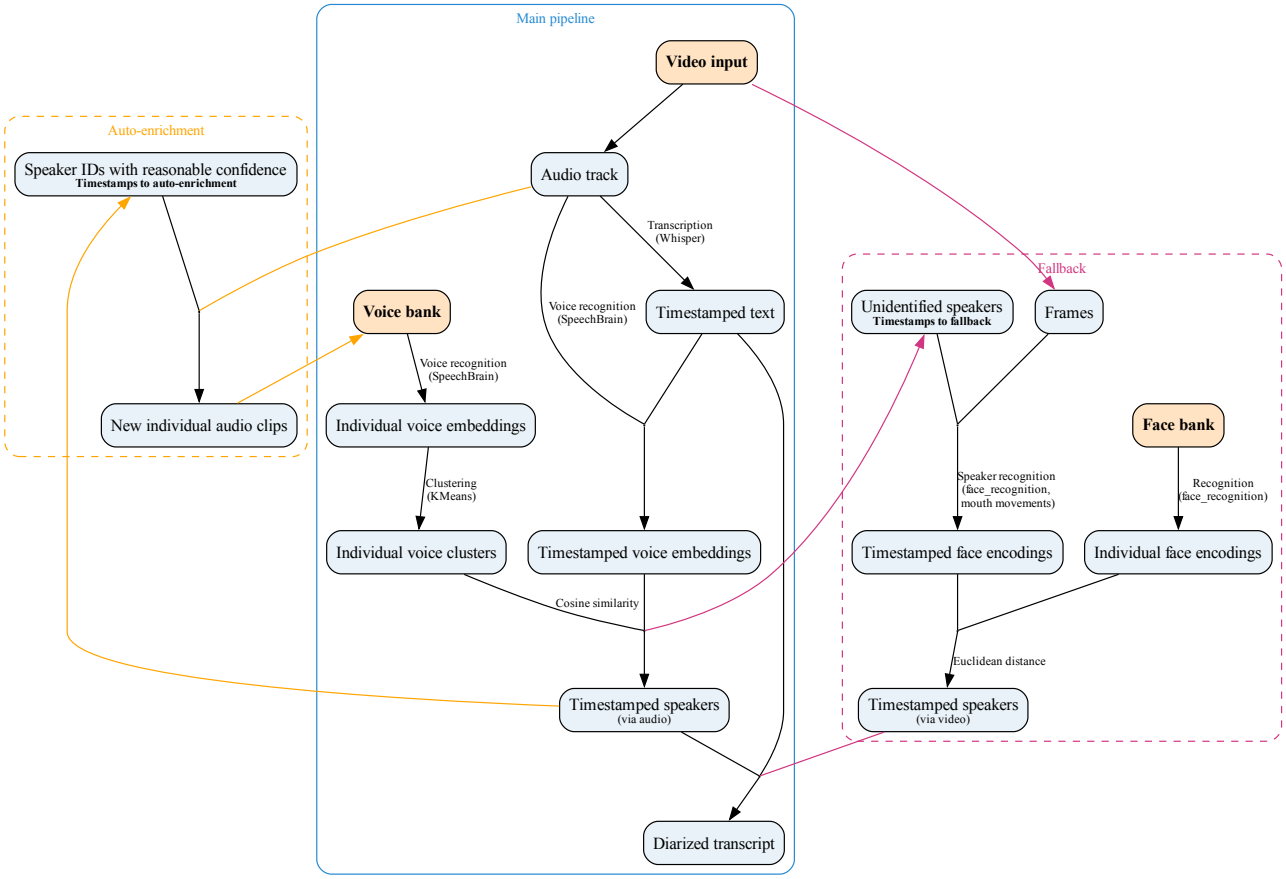
5 Self-learning applications and extensions

The pipeline that the previous section detailed is relatively simple to put into practice and performs well in accuracy metrics, as section 6 will show. One relevant flaw, however, is that it requires significant preparation work by using manually-collected face and voice banks. While faces are relatively quick to collect, audio clip choice and trimming is more time-intensive, especially if one takes care to source clips from different circumstances. It is also evident that speaker identification through video is less accurate and more computationally intensive than its audio counterpart. In this section, I tackle this by proposing two extensions of the baseline pipeline that automate previously-manual processes.

Firstly, I implement a self-learning audio step, which aims to improve audio detection to a point that eventually obviates the need for the video step, improving accuracy and reducing runtime. Secondly, I propose an auxiliary function that automates the collection of voice clips, so that the user only needs to manually compile the face bank. These significantly reduce the required preparation work the pipeline requires, facilitate their expansion to new datasets (also from new environments or different languages), and create a function that becomes more accurate the more it observes each speaker.

To finish, this section also introduces an add-on to the pipeline that uses a GPT model to infer speaker identities for particularly challenging audio segments.

Figure 3: Self-enriching pipeline



Notes: Diagram representation of the self-enriching pipeline. The fallback branch only runs when a segment cannot be confidently attributed to a known speaker via the main branch, and only for those segments. The voice bank enrichment step runs after the diarization process and only benefits subsequent applications.

5.1 Auto-enriching voice bank

The self-learning extension is a relatively straightforward add-on to the baseline pipeline and is illustrated by figure 3. Most of the process runs exactly as it did before, with two relevant changes: the self-learning module itself and the use of several centroids (K-Means) for each speaker representation.

5.1.1 Self-enrichment

The first of the two main changes this pipeline introduces is the self-learning module, which is mostly implemented by the *auto_enrich_voicebank* function, which runs after the timestamped speakers via audio phase and can be found in full in appendix A7.

At this stage, the pipeline checks whether there are any audio segments of at least a given duration (*auto_ref_min_dur*) that were attributed to a given speaker within a confidence interval (*auto_ref_min_conf* to *auto_ref_max_conf*). It also checks for consecutive segments that, together, meet *auto_ref_min_dur* with an average confidence within the set interval. Finally, it considers how many voice clips are already available for the speaker in question in the voice bank, proceeding with the collection only if that number is lower than *auto_ref_max_clips*.

auto_ref_min_dur is meant to guide the function into collecting a lower number of richer segments. By

default it is set to 10 seconds, which ensures each clip has a relatively long window of actual speech and prevents the proliferation of short, minimally-informative segments. It also significantly lowers the chance of speaker misidentification, as it is very unlikely that a misattribution achieves high confidence for such a long segment. Finally, it ensures that occasional interruptions by other speakers (which are necessarily short, otherwise they would have broken up the current segment) are rare when compared to the main individual's speaking time, introducing some welcome variability in how they sound without significantly skewing their vector representation.

auto_ref_min_conf and *auto_ref_max_conf*, on the other hand, ensure that the collected segments are not too dissimilar or too similar to the speaker's current representation, respectively. The former is set to a default value of 0.65 and meant to prevent wrong attributions – at that value or above, we can be very confident that we are only picking up the intended speaker. The latter defaults to 0.8 and stops the module from collecting segments that are functionally almost equivalent to the ones we already have in the voice bank – if a given segment is already very confidently identified by the current speaker profile, adding it to the index will provide only marginal improvements (and can even be harmful if over-fitting on usual circumstances pushes more-abnormal instances to outlier status). Through these checks, the focus of this collection step is on building a richer representation of each speaker.

Finally, *auto_ref_max_clips* limits how many clips the voice bank can contain. By default, I have set it to 100, but this is arbitrary and depends on the application. With this value, a balance is struck between building a sizable dataset, occupying a relatively low amount of disk space (after running the self-learning step for my entire dataset, I obtained a total of 3944 clips occupying 4.69 GB of space), and minimizing runtime once all speakers have reached their limit.

This collection is attempted whenever a new video is diarized, meaning care must be taken not to run the same video twice with this feature enabled (a *self_enrich=True/False* switch is implemented to account for this). After each collection step, a new index is created that accounts for the new voice bank composition and makes it available for the following diarizations.

5.1.2 K-Means clustering

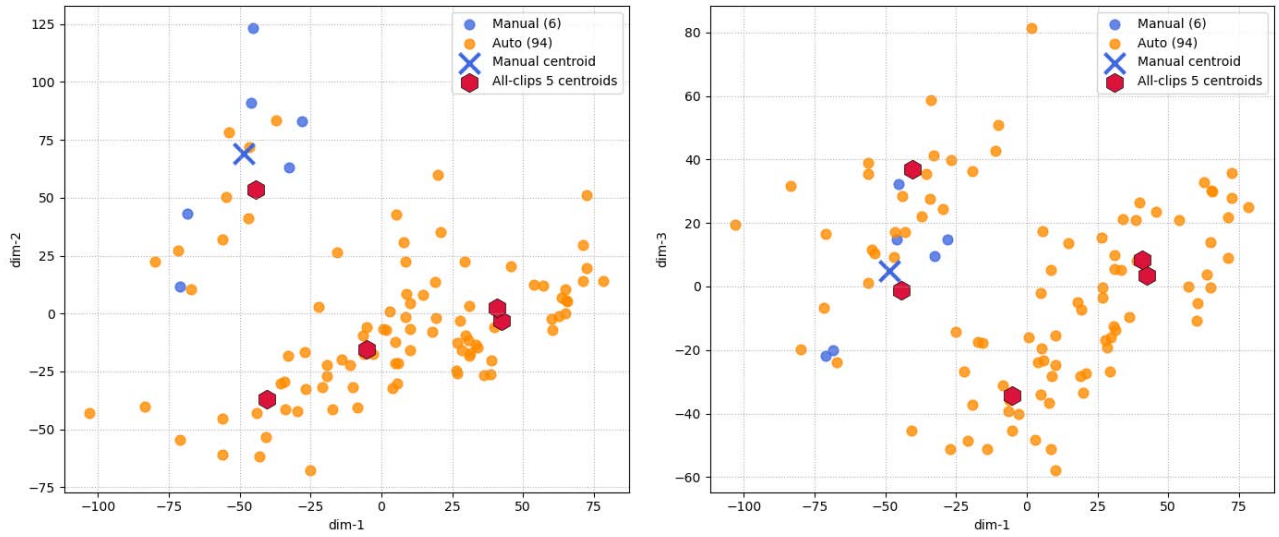
The baseline pipeline took a simple approach to each individual's audio profile: it aggregated the six initial voice samples into a single centroid by taking each dimension's mean. This approach is a good fit for a small amount of samples as it can average out artifacts, momentary abnormalities or temporary conditions (e.g. speaker illness, microphone malfunction, interruption by other speakers). The large amount of voice samples per individual that the self-learning method allows, combined with the within-speaker variety encouraging collection method, however, suggest a more complex method of representing audio profiles. As such, this pipeline also implements K-Means clustering per speaker.

Briefly, this method consists of aggregating the N samples into a K number of centroids, rather than a single one. To determine K, I use the following rule:

$$K = \begin{cases} 1, & N < 10 \\ 2, & 10 \leq N < 25 \\ 3, & 25 \leq N < 50 \\ 4, & 50 \leq N < 100 \\ 5, & N \geq 100 \end{cases}$$

When less than 10 samples are available, the pipeline reverts to a single centroid, which is equivalent to the baseline application. As the amount of samples grows, so do the number of centroids. A centroid can be thought of as a speaker state that impacts how they sound like: healthy, ill, young, old, interrupted by another person, using a different kind of microphone. The rule attempts to balance parsimony with an effective use of the increased

Figure 4: Manual vs. automatic speaker centroids: João Cotrim de Figueiredo



Notes: PCA(2)-obtained depiction of centroids before and after a full self-learning process. Each circle represents the audio embedding of a potential speaker's single voice bank clip, with blue circles referring to those initially collected by hand and orange circles to those collected automatically. The blue "X" marks the single centroid obtained only from the manually-collected clips, while the red hexagons mark the five centroids obtained from all 100 clips (manually- and automatically-collected).

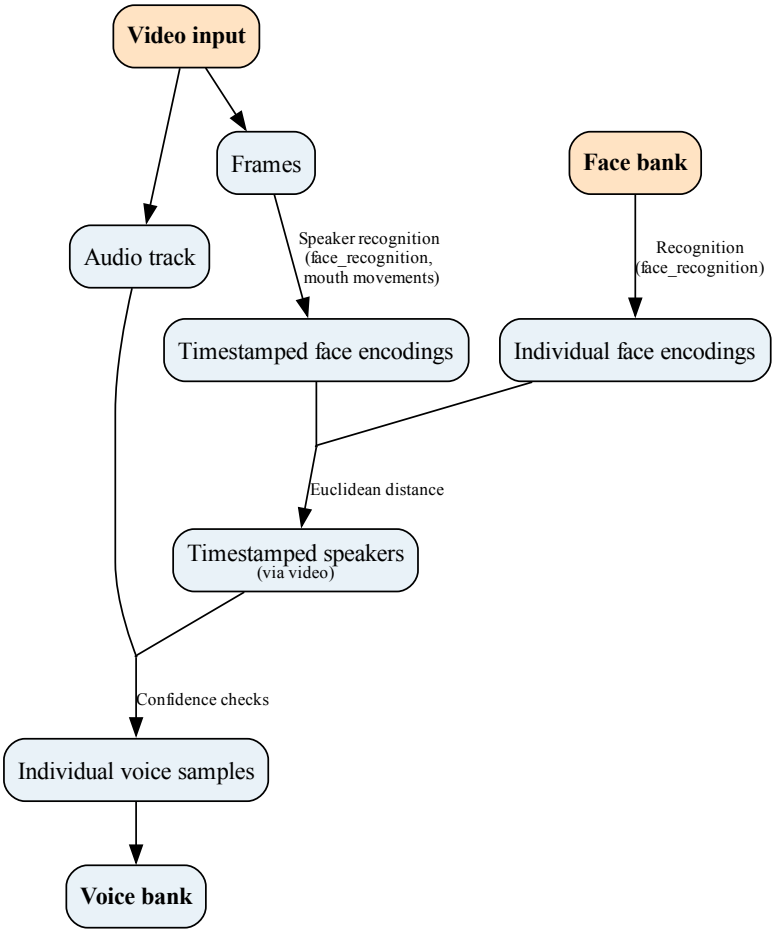
amount of available data – too many centroids for too few samples risk taking an aberration for a state, and too few centroids risk aggregating two real states. Further note that this model will often have an unbalanced voice bank, especially while in the initial stages of usage and when meeting new individuals. This approach also allows for the graceful handling of this difference in N between speakers, keeping an adequate number of centroids for each.

Obviously, these values are arbitrary and tailored to my use case; different applications might call for different thresholds. To tweak them, one should bear in mind that, while a higher N can call for higher K, there should be an upper bound to K that is grounded on reality – each person does not sound an infinite number of ways, but only a few. Increasing K also makes this pipeline more computationally intensive, and increasing N to support it requires more storage space. Figure 4 displays two graphical representations of this process, using principal component analysis to reduce dimensionality to a representable 2-D. With this technique, dimension 1 is the principal component that explains the most variance, while dimension 2 explains the second-most variance. Given that the actual speaker embeddings are 192-D, proximity is imperfectly depicted (nonetheless, it is as accurate as it can be in the 2-D space). A dim-1 vs. dim-3 plot (which, naturally, explains the third-most variance) is also included to provide additional perspective.

The figure, which illustrates the specific case of Iniciativa Liberal MEP and former-leader João Cotrim de Figueiredo,[17] exemplifies the gains from implementing the self-learning module and allowing for additional centroids: firstly, the full-dataset centroid spread suggests this speaker has two to four states. Secondly, the centroid obtained from only the manually-collected clips seems to somewhat overlap with one of the five centroids obtained from the full sample – by employing only that measure, one would not account for the majority of the speaker's states. Finally, centroids seem sufficiently separated for the use of a single value (K=1) to entail relevant data loss.

[17] https://www.europarl.europa.eu/meps/en/257057/JOAO_COTRIM+DE+FIGUEIREDO/home.

Figure 5: Automatic voice collection pipeline



Notes: Diagram representation of the automatic voice collection pipeline. This process is meant to be run before the main pipeline, in place of manual voice sample collection.

5.2 Independent voice collection from face bank

While the previous self-enriching pipeline provides a means to build a large voice bank with only a few initial clips, it still requires manual collection for that initial set. Manually collecting voice clips from a video is a time-consuming process that requires the extraction of the audio track from the video, watching the video itself to identify suitable clips for harvesting, and the use of specific software such as Audacity to trim and export samples. This section provides an auxiliary pipeline that automates this process so that only the manual collection of pictures – a significantly faster process – is required. Figure 5 presents a diagram depiction of this module.

This pipeline is meant to be run before the main pipeline and works similarly to the video fallback. Unlike the baseline, which required a video input, a face bank and a voice bank, this function takes only the first two as inputs. It first checks the people index and identifies those to be sought in the video through a manual "auto-collect" flag, whitelisting individuals who are so flagged and have at least one entry in the face bank. It then analyzes every frame in the video input with the face_recognition package, matching identified faces to the whitelist if they meet a Euclidean distance threshold – here I am using a slightly more stringent maximum of 0.45 than in the video fallback to minimize identification errors.

Using a similar method of active speaker detection as in the video fallback, the function then identifies segments of at least 10 seconds where the same person is identified as speaking. For all matches, the corresponding audio clip is collected and added to the voice bank.

Requiring a speaker to be for at least 10 seconds on camera and speaking is a stringent requirement that makes it virtually impossible for the expected speaker not to be actually speaking, as any relevant interruption would warrant a shot of another speaker or the moderator. To test this method, I apply it to the multi-party debates that were excluded from the main dataset,[18] for which no manual voice collection was made, and find that it returns 100% speaker identification and collection accuracy. One must bear in mind, however, that fully analyzing videos is computationally demanding – this method saves manual work effectively, but requires time and computing power to run.

5.3 GPT correction

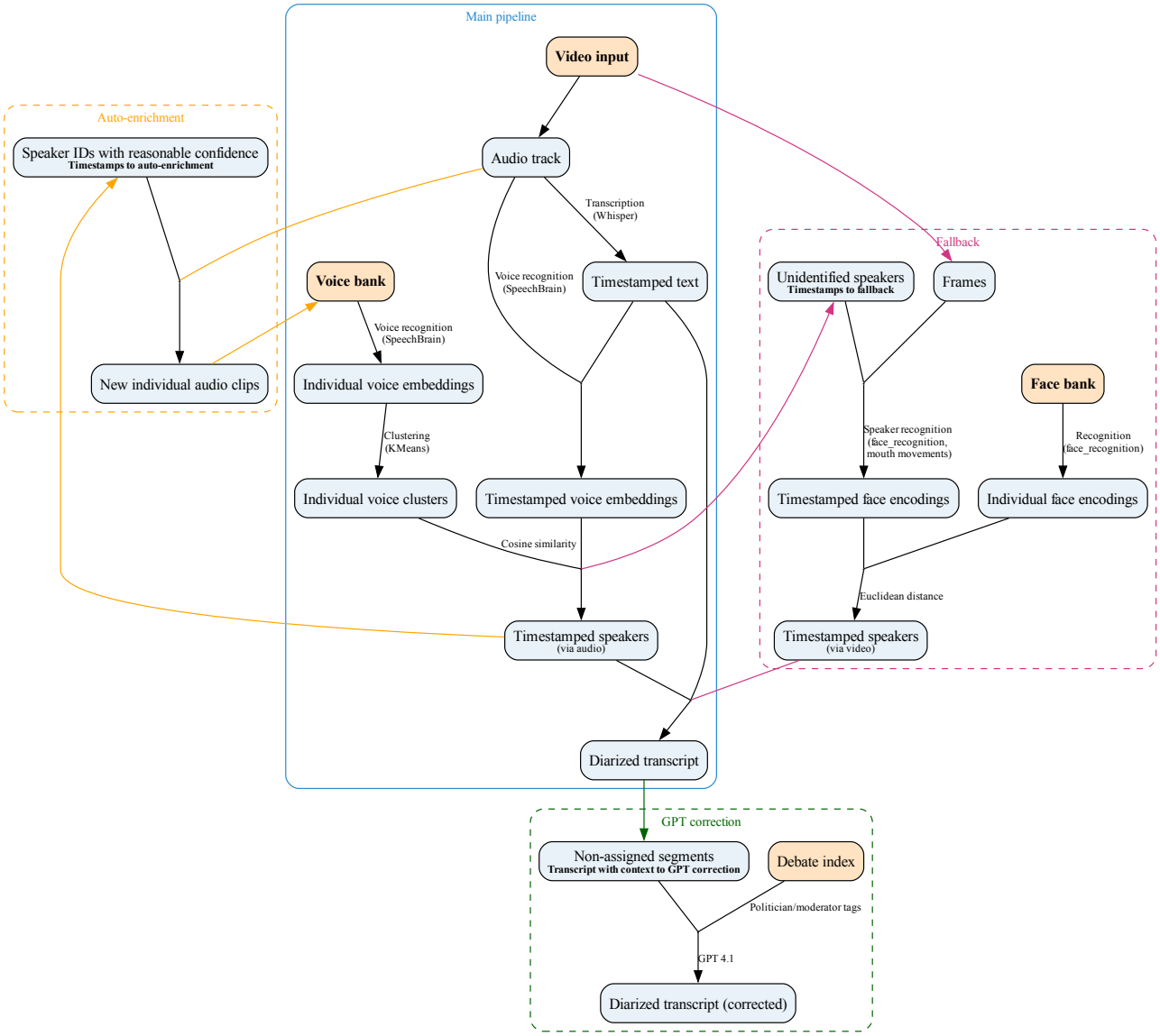
While the implementation of these expansions yields accuracy gains, there still remain segments that neither the audio nor video step could assign to a known speaker. They are, however, now surrounded by diarized segments that provide rich context and should make inferring missing speakers relatively easy. In this sub-section, I leverage this by implementing an LLM-powered inference step for these segments – a relatively straightforward process that can be run on the previously-obtained diarized transcripts. I have chosen to only undertake this step for "Unknown" labels, but a possible extension would be to employ it also for low-accuracy audio or video identifications as a verification step. Figure 6 graphically describes this process and how it fits into the wider diarization pipeline.

In this step, each diarized transcript is analyzed in turn. To begin, all speakers in the transcript are classified as either politician or journalist through the debates index (illustrated in table 1). Journalists are assumed to be moderating and party members to be debating, which influences their tone and spoken content and as such helps infer who delivered a given line. Afterward, each "Unknown" segment is passed to GPT (in this case, GPT-4.1), with a context window of the eight surrounding segments (four before and four after), according to the function presented in appendix A8. A specific example of a passed prompt is as follows:

> You are an AI assistant helping to attribute debate turns.
> 1. Focus on the line that contains ###TARGET###.
> 2. Choose which candidate is speaking in that line.
> 3. If <70 % confident, reply "Unknown".
> CANDIDATES:
> - André Ventura (politician, Partido="CH")
> - Catarina Martins (politician, Partido="BE")
> - Rosa Oliveira Pinto (moderator, Partido="jorn")
> CONTEXT:
> André Ventura: Tudo farei para retirar António Costa do poder,
> André Ventura: para devolver o poder de compra que estes senhores tiraram aos portugueses
> André Ventura: e para conseguir fazer uma coisa que não foi feita.
> André Ventura: É acabar com este país em que metade trabalha para sustentar a outra metade.
> Unknown: ###TARGET### É isso que temos que acabar.
> Rosa Oliveira Pinto: Há um outro cenário de que temos que falar, Catarina Martins.
> Rosa Oliveira Pinto: Em caso de governo sem maioria do PSD,
> Rosa Oliveira Pinto: preferem abster-se, viabilizando orçamentos,
> Rosa Oliveira Pinto: ou preferem deixar o PSD a negociar com o André Ventura?

[18] Described in section 3.1.

Figure 6: Full pipeline



Notes: Diagram representation of the full pipeline. The fallback branch only runs when a segment cannot be confidently attributed to a known speaker via the main branch, and only for those segments. The voice bank enrichment step runs after the diarization process and only benefits subsequent applications. GPT correction is applied ex-post.

Respond with a single JSON line:
{"speaker": "<name or Unknown>", "confidence": <0-1>, "reason": "<short>"}

GPT is further instructed to implement a certainty filter through "at least 70% accuracy". This value does not refer to anything specific and one should not expect it to be followed according to any consistent or grounded rule; it is simply a way to direct GPT to only make informed guesses regarding who is speaking. I also instruct it to output directly in JSON format, which it does competently and consistently. Applied to my 108-debate, approximately 60h dataset, this process was relatively quick and cheap – around 40 minutes and $3 for all segments that remained "Unknown" after running the full pipeline. As the following section will show, it contributes relevantly to the accuracy of the entire process.

6 Accuracy reports

This section presents some accuracy metrics for these pipelines. To establish a ground truth, I manually label speakers in six of the transcripts yielded by Whisper (one from each covered election, totaling approximately 203 minutes). This means the evaluation set is not a perfect benchmark – occasionally, speaker overlap means that not all spoken lines are accounted for. While this prevents me from presenting more elaborate statistics like the diarization error rate[19] (as in Sharma and Narayanan, 2022 and Mingote et al., 2024, for example), it still allows for comparison between instances and quantification of the gains from self-enrichment and the GPT correction. For each debate, I present metrics for the baseline, self-learning method after learning from the entire dataset, and final, GPT-corrected transcripts.

Table 3 describes this exercise, with statistics for each of the manually-labelled debates. It details the total number of identified spoken segments; "substantial segments" – segments that are at least three words long – which are more likely to carry a message; and (V) segments, referring to segments that were attributed to a speaker via the fallback video pipeline. All other segments ($segments - (V)\,segments$) are (A) segments, attributed through the main audio pipeline. From these, I compute the following statistics:

$$
\begin{aligned}
Overall\,accuracy &= \frac{correctly\,attributed\,segments}{all\,segments} \\
Substantial\,accuracy &= \frac{correctly\,attributed\,subst.\,segments}{subst.\,segments} \\
(A)\,accuracy &= \frac{correctly\,attributed\,(A)\,segments}{(A)\,segments} \\
(V)\,accuracy &= \frac{correctly\,attributed\,(V)\,segments}{(V)\,segments} \\
Time\,error\,rate &= \frac{time\,length\,of\,correctly\,attributed\,segments}{total\,spoken\,time} \\
Word\,error\,rate &= \frac{word\,length\,of\,correctly\,attributed\,segments}{total\,spoken\,words}
\end{aligned}
$$

Figures are presented for both baseline, self-enriched, and GPT-corrected sets. The baseline stats correspond to diarizing these six videos using only the manually-collected face and voice banks. For the self-enriched stats, the self-learning pipeline was run for my entire debate set and then run again for these six instances. Not all speakers were fully enriched (i.e. until 100 voice samples were reached) – figure 7 details exact numbers for each. These results were then passed on to the GPT correction process for the corrected set.

Several conclusions can be taken from this exercise. Firstly, diarization accuracy is respectable even without any training, presenting an overall accuracy of 93.76% across all six videos and suggesting that, even without training, the pipeline can perform to a high standard. Self-enrichment, however, elevates this figure to 96.31%, with time and word error rates of around 2%, showing that the self-learning process is fruitful and yields a highly reliable diarization method. All metrics improve when compared with the baseline save for (V) accuracy, but this should not be taken as worse performance – it is simply the case that, after self-enrichment, the audio step becomes more

[19] Fiscus et al., 2006.

Table 3: Accuracy report

Baseline

Election	Parties	Segments	Subst. segments	(V) segments	Overall acc.	Subst. acc.	(A) acc.	(V) acc.	TER	WER
2019 Eur.	CDS-PP x BE	756	733	56	94.18%	94.95%	96.14	69.64	3.26%	3.66%
2024 Eur.	AD x IL x PS x L	1414	1380	63	91.02%	91.81%	92.82%	52.38%	5.67%	5.65%
2019 Leg.	PS x CDU	772	719	11	97.93%	98.75%	98.03%	90.91%	0.98%	1.04%
2022 Leg.	BE x CH	667	654	13	95.05%	96.18%	95.72%	61.54%	2.52%	2.76%
2024 Leg.	PS x IL	672	656	8	97.02%	97.71%	97.89%	25.00%	1.48%	1.66%
2021 Pres.	Ind. x IL	862	813	69	90.60%	92.00%	92.18%	72.46%	6.22%	6.19%
		5143	4955	220	93.76%	94.67%	95.06%	64.54%	3.6%	3.75%

Self-enriched

Election	Parties	Segments	Subst. segments	(V) segments	Overall acc.	Subst. acc.	(A) acc.	(V) acc.	TER	WER
2019 Eur.	CDS-PP x BE	756	733	26	96.16%	96.73%	97.53%	57.69%	2.22%	2.24%
2024 Eur.	AD x IL x PS x L	1414	1380	34	94.41%	94.86%	95.80%	38.24%	3.48%	3.49%
2019 Leg.	PS x CDU	772	719	8	98.32%	99.03%	98.56%	75.00%	0.86%	0.86%
2022 Leg.	BE x CH	667	654	8	96.40%	97.55%	96.97%	50.00%	1.77%	1.61%
2024 Leg.	PS x IL	672	656	5	98.07%	98.78%	98.65%	20.00%	0.54%	0.85%
2021 Pres.	Ind. x IL	862	813	17	96.29%	97.05%	96.57%	82.35%	2.40%	2.31%
		5143	4955	98	96.31%	96.98%	97.13%	54.08%	2.05%	2.10%

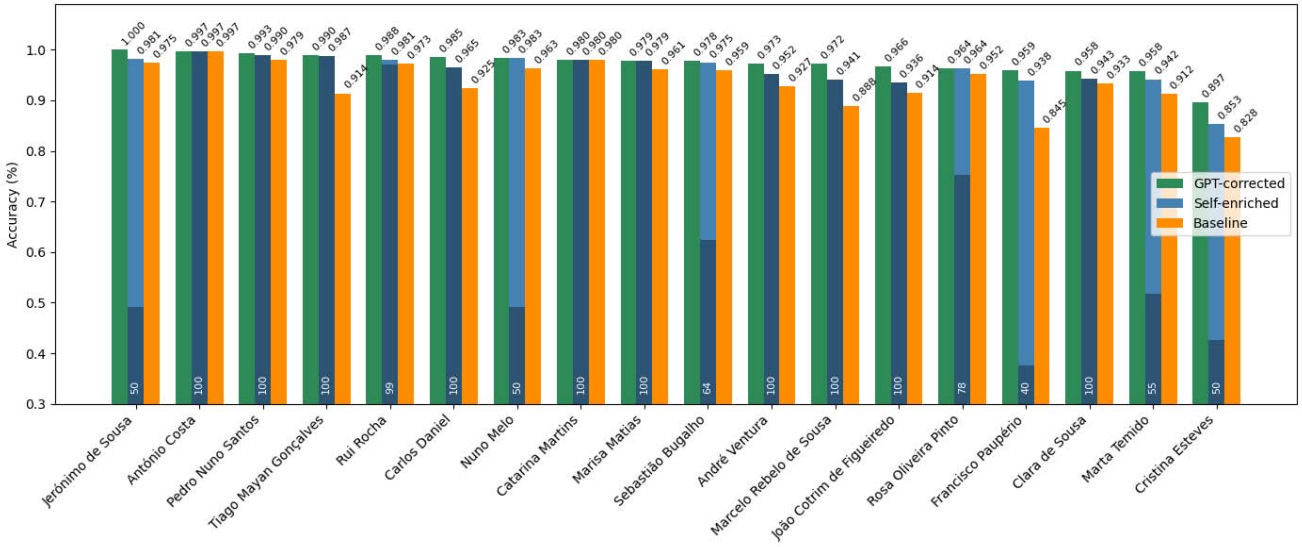
GPT-corrected

Election	Parties	Segments	Subst. segments	(V) segments	Overall acc.	Subst. acc.	(A) acc.	(V) acc.	TER	WER
2019 Eur.	CDS-PP x BE	756	733	26	96.83%	97.00%	98.22%	57.69%	1.92%	2.02%
2024 Eur.	AD x IL x PS x L	1414	1380	34	96.25%	96.52%	97.68%	38.24%	2.65%	2.48%
2019 Leg.	PS x CDU	772	719	8	99.35%	99.58%	99.61%	75.00%	0.40%	0.35%
2022 Leg.	BE x CH	667	654	8	97.45%	98.47%	98.03%	50.00%	1.42%	1.13%
2024 Leg.	PS x IL	672	656	5	98.66%	99.09%	99.25%	20.00%	0.42%	0.67%
2021 Pres.	Ind. x IL	862	813	17	97.80%	97.91%	98.11%	82.35%	1.86%	1.82%
		5143	4955	98	97.53%	97.86%	98.38%	54.08%	1.57%	1.55%

competent in identifying speakers and thus leaves only the most difficult cases for the video step (less than half those in the baseline, in this specific case), degrading its performance metrics. Finally, GPT correction provides further accuracy improvements across the board of slightly above one percentual point, bringing overall accuracy to a final figure of 97.53%.

Figure 7 presents overall accuracy statistics in this set broken down by speaker, rather than debate. The orange bars refer to accuracy obtained through the baseline process, blue bars refer to accuracy after self-enrichment and green bars to accuracy after GPT correction. The degree to which the blue bar is filled and the bottom label represent the percentage of training completed – how many out of 100 voice clips were collected by running the code on my entire set of debates. Three main remarks can be made on it: firstly, the self-learning process appears to improve performance, but not equally for all speakers – two do not seem to benefit at all (António Costa and Catarina Martins), while the others display gains that range from negligible to sizable. Secondly, a full training set is not (always) necessary to significantly improve diarization accuracy – the largest observed improvement from baseline to self-enriched happens with Francisco Paupério, who has the least voice clips in his voice bank. Thirdly, the GPT correction process frequently improves performance (sometimes sizably, sometimes minimally) and never

Figure 7: Overall accuracy by speaker, baseline and self-enriched



Notes: Overall diarization accuracy per speaker in the ground truth set (six debates), baseline vs. self-enriched vs. GPT-corrected. The self-enriched bars' dark filling and bottom label correspond to the percentage of completed self-training.

harms it, proving to be a beneficial addition.

These observations can show that different speakers require different amounts of training – some speakers, due to a more recognizable manner of speech or more singular voice characteristics, are likely easier to identify. One must bear in mind, however, that these results stem from a debate per speaker (save for Clara de Sousa, who moderates three of the debates in this ground truth set), and as such might not be representative of the entire dataset. António Costa's case is paradigmatic: here we are looking at the September 2nd 2019 António Costa x Jerónimo de Sousa debate – one of the most civil in the entire dataset with, as such, minimal speaker overlap. These conditions maximize the performance of diarization and definitely do not occur in all debates.

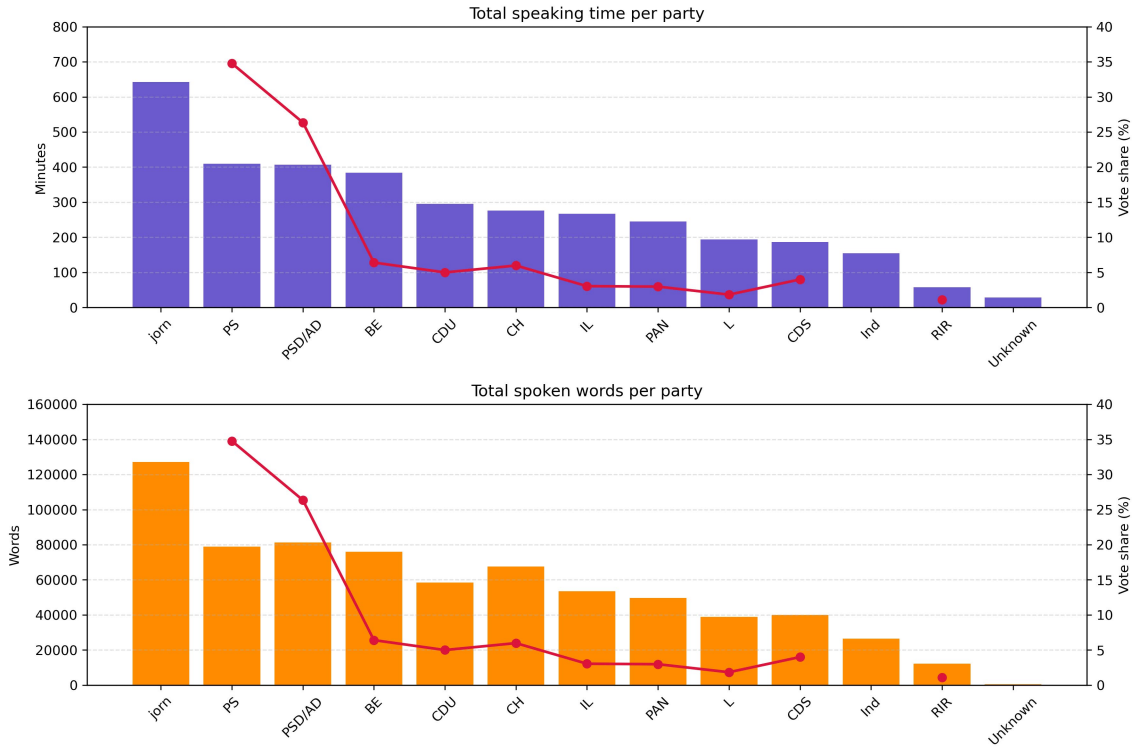
Finally, one might conclude from the figure that 100 voice clips is an arbitrary amount and might either be overkill or not enough to maximize accuracy gains. This is fair criticism and relatively easy, if computationally intensive, to substantiate – a further research avenue is to thoroughly test the progressive accuracy gains of increasing the voice bank for the same ground truth set, which should be paired with a study of the ideal number of centroids.

7 An application - policy proposal tallying

Having built the full pipeline, the natural next step is to provide an example application. In this section, I demonstrate how to perform a simple analysis of the obtained transcripts through the tallying of policy proposals from a corpus of campaign debates. To this end, I will be using the set of debate transcriptions that I previously obtained and GPT-4.1 to interpret them.

I begin with figure 8, a graphical depiction of the distribution of speech time across parties compared with their overall electoral results in the covered elections (note that journalists appear over-represented because they are present in all debates, rather than a subset). The figure suggests that the speaking opportunities parties receive do not depend solely on their past or expected results, which exemplifies a question that this dataset could help explore. Note that this plot also roughly illustrates the percentage of segments that remain labelled as "Unknown" after the GPT-correction process. The discrepancy between the time and number of words in this figure is mostly due to a Whisper-related phenomenon that occasionally labels the period between the beginning of a recording and the first person speaking, or between the last person speaking and the end of the recording, as an empty "Unknown"

Figure 8: Representation per party



Notes: Total spoken minutes and words per party in the entire dataset (bars, left y axes) and average electoral result in percentage (red line, right y axes) per represented party. "Ind" refers to the independent presidential runs by Marcelo Rebelo de Sousa (originally PSD) and Ana Gomes (originally PS), which are excluded. The Chega average vote includes the 2019 European elections Basta coalition result.

segment. These segments have no impact on performance and should be ignored.

7.1 Policy proposal recognition

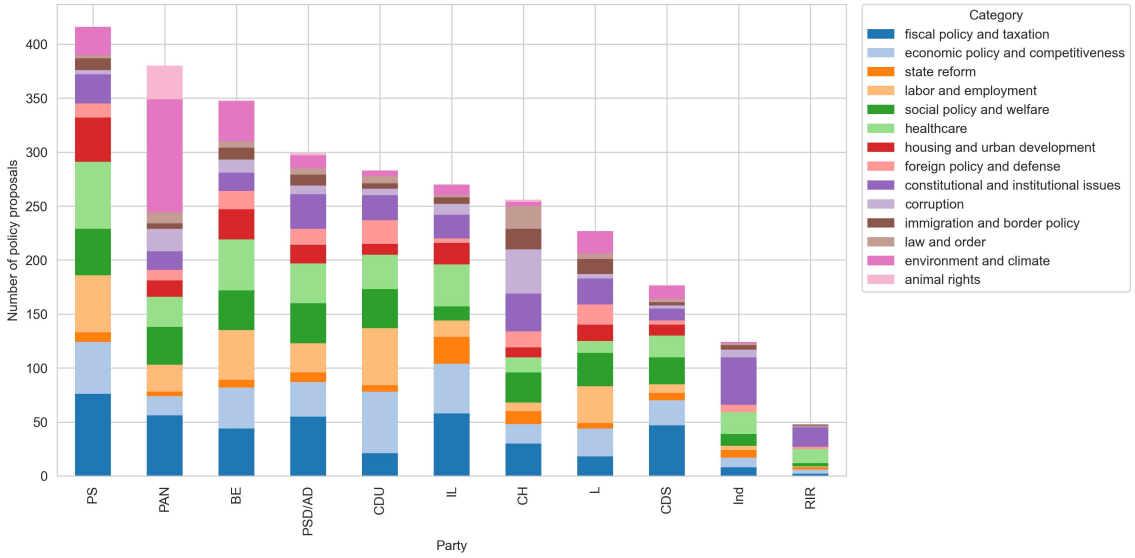
In addition to the more straightforward applications like the one we have just seen, this dataset can be used to perform more nuanced analysis, especially with the help of LLM-based approaches. Over this sub-section, I propose and design a function that uses GPT-4.1 to collect all policy proposals that are made in each debate, aggregates them into separate policy fields, and uses the obtained tallies to describe the policy content of each party's discourse across the different debates they participate in and the different opponents they face.

I begin by running a loop that takes a list of speakers and, for each, collects all segments uttered by them in each debate in the dataset. It then feeds each debate's concatenated segments to GPT in turn, according to the function in appendix A9. This, for each debate-speaker pair, creates a text file with each found instance of a policy proposal and an exact quote as evidence.[20] Once this is done, I pass those text files to GPT once more and in turn, to join duplicate proposals and group them into policy categories.

Note that it would be possible to do this task together with the previous one. Separation, however, lowers the likelihood of content being overlooked, which is especially relevant when a counting request is made. I arbitrarily choose the following 14 policy categories from observing the debate transcripts, which can be tailored to each application: fiscal policy and taxation, economic policy and competitiveness, state reform, labor and employment, social policy and welfare, healthcare, housing and urban development, foreign policy and defense, constitutional and institutional issues, corruption, immigration and border policy, law and order, environment and climate, and

[20] Meaning that, for each debate, as many files as speakers, excluding the moderator, are created.

Figure 9: Policy proposals by party and category



Notes: Total sum of all policy proposals made by each party's representatives across the full dataset.

animal rights. Appendix A10 presents the proposal classification function.

In the end, this section outputs a .csv table with the total and per-category policy proposal tally per debate – speaker pair. Finally, using the debate index illustrated in table 1, the table is enriched with each speaker's opponent, the party affiliation for both, the election the debate occurred in, and who the moderator was. This allows for some interesting analysis that the next sub-section will illustrate.

7.2 Policy proposal analysis

I now turn to the graphical depiction and discussion of the policy proposal dataset we just generated. Figure 9 presents each party's policy proposal distribution by category in a stacked bar format. Six elections are accounted for: 2019 and 2024 European elections, 2019, 2022 and 2024 legislative elections, and 2021 presidential elections. It shows substantial differences among parties, not only in the quantity of proposals but also in their category mix. Note, for example, how PAN comes second even though they are one of the parties that obtained the lowest average vote share in this timeframe, and how they focus significantly more on the environment and animal issues than other parties; how the liberals focus the most on state reform; or how Chega make the most proposals on corruption and constitutional matters. The observations for independent and RIR candidates refer only to the 2021 presidential elections (Marcelo Rebelo de Sousa and Ana Gomes for the former and Vitorino Silva for the latter): their focus on constitutional rights is due to the nature of the president's role, and on healthcare is due to the COVID-19 pandemic that was then being felt.

An alternative way to look at this data is to break it down by election. Figure 10 illustrates this exercise, showing the evolution of salient issues: fiscal and economic policy were the most salient in the 2019 European elections, whereas by the 2024 European elections, amidst the war in Ukraine and increasing immigration in Portugal, the discussion was focused on immigration and defense. The advent of the current Portuguese housing crisis can also be identified in the attention spike that topic receives in the 2024 legislative elections.

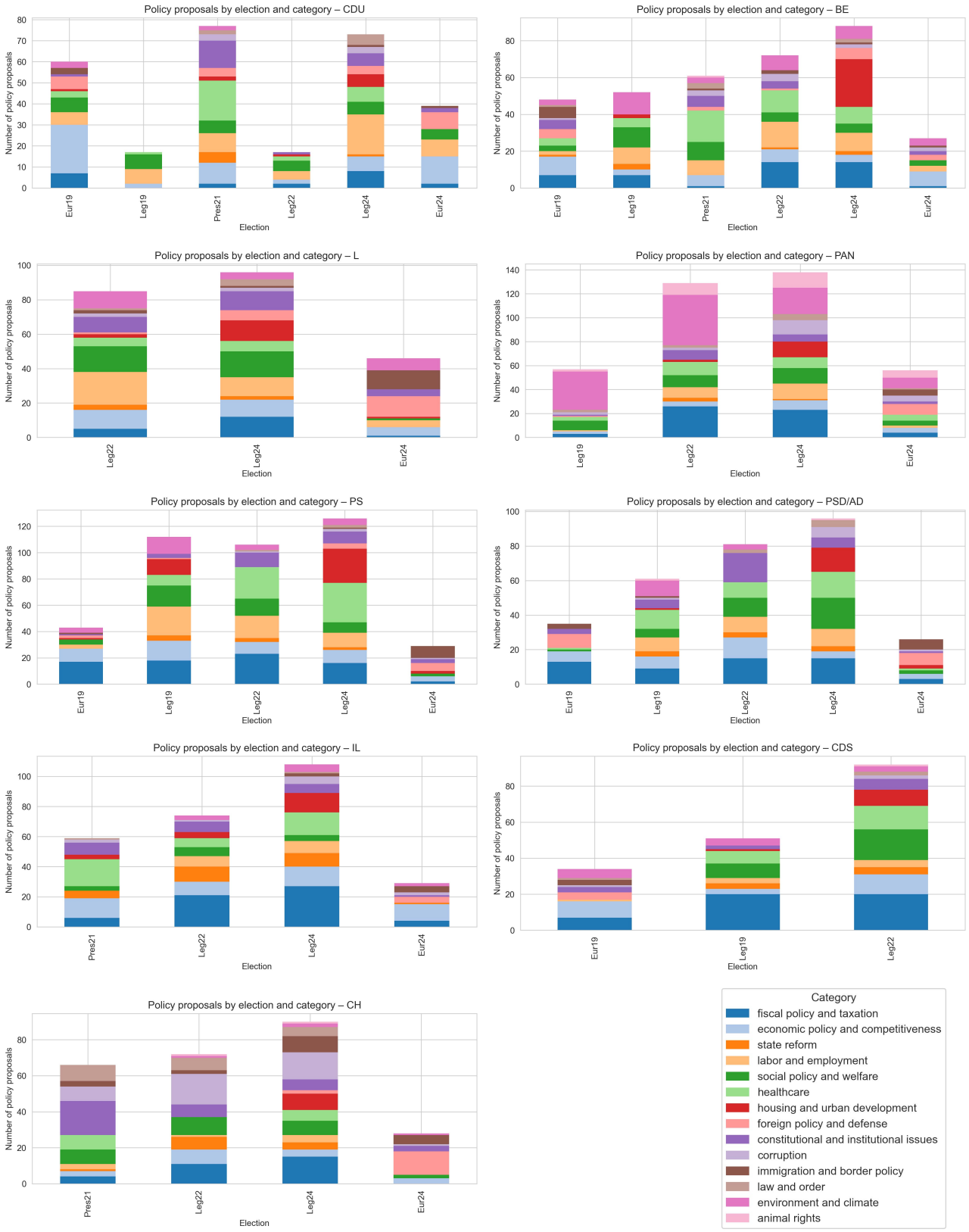
Yet another possible visualization of this dataset is by party, across elections. Figure 11 presents this approach, where we can identify, for example, the evolution of Chega's message (in particular, an early relatively low focus on immigration), as well as PAN's attempt to move away from a single-issue party behavior.

We can also analyze each party's message composition according to their opposition, which might shed light

Figure 10: Policy proposals by party and category (breakdown by election)

Notes: Total sum of all policy proposals made by each party's representatives across each election.

Figure 11: Policy proposals by election and category (breakdown by party)



Notes: Total sum of all policy proposals made by each party's representatives across each election.

on topic matching/dismissal strategies.[21] Figure 12 does exactly this and allows us to quantify a few interesting observations. One example is how Chega takes an apparently similar approach to debating with BE and IL – in principle quite different opponents but with some contact points regarding social/cultural liberalism. One can also see how all parties but CDU, CDS and Chega – the socially conservative parties – engage relevantly with environmental topics when debating PAN.

Finally, we can use this data to compute party proximity. Given this exercise only quantifies proposals, we can only evaluate and compare how much attention each party dedicates to each topic. However, it would be a relatively simple extension to quantify the directionality of policy proposals, allowing for analysis of proposal, rather than just focus proximity. To this end, figure 13 presents a reduction to two dimensions using principal component analysis across all elections. Figure 14, on the other hand, presents individual PCA reductions for each election.[22] In general, the right/left-wing split is visible, showing they put emphasis on a different combination of policy areas; one can also see how CDU and Chega, as arguably the further-left and further-right parties, respectively, are always distant from other parties; and finally it is also interesting to see the progressive coming apart of PS and PSD/AD across time in legislative elections.

8 Concluding remarks

This work lays the blueprint for a complete method for obtaining diarized transcripts from videos, especially tailored to electoral debates and similar kinds of media (such as politician one-on-one interviews). It goes on to provide accuracy metrics and exemplify a few possible applications, such as an analysis of preferred policy topics per party across different electoral campaigns. Electoral debates, as an input, are a particularly rich source of information, shedding light on topics such as politician policy proposals, debate style, degree of civility, consistency and evolution. Through challenges by opponents and moderators, they also have the potential to reveal information that might remain unsaid or unnoticed in other media. The accuracy of the method I propose – over 97% – means it can be used with near-irrelevant loss of data for most applications.

This is a relevant pursuit for several reasons. Firstly, it opens the door for the analysis of virtually any debate for which there is a video record, rather than only those for which official transcripts are made available, massively increasing the potential amount of data for research focused on this kind of media. Secondly, it easily enables cross-country analysis of media without any additional work – a simple true/false flag dictates whether transcripts are delivered in English or their original language. Once more, this unlocks the door to new and larger-scale potential research angles. Thirdly, the method I present fits firmly in with the advancement of political science towards more objective and replicable inputs, which is, for scientific pursuits, a goal in itself that the literature often discusses.

Additionally, the pipelines I propose are modular, mostly open-source (with GPT as the sole exception) and easy to deploy separately or adapt to different objectives or inputs. As such, it would be simple to repurpose this code for inputs with different characteristics (e.g. debates or interviews with more speakers, shorter format-videos such as those found on social media, broadcast news), each with equal potential to expand the available amount of data for analysis. The advance of LLM technology further compounds on all these factors by making this kind of data more useful – it has never been easier to turn political text into quantitative aggregates.

Obviously, this work is not pioneering methods-wise, as the existing literature on speaker diarization can attest to. However, by bringing the method closer to political science, through easily adaptable code and relevant example usages, I hope to bridge a gap that can deliver sizable future gains. There has never been as much available data as there is today – maximizing how much can be extracted from it is a pursuit with immense potential returns.

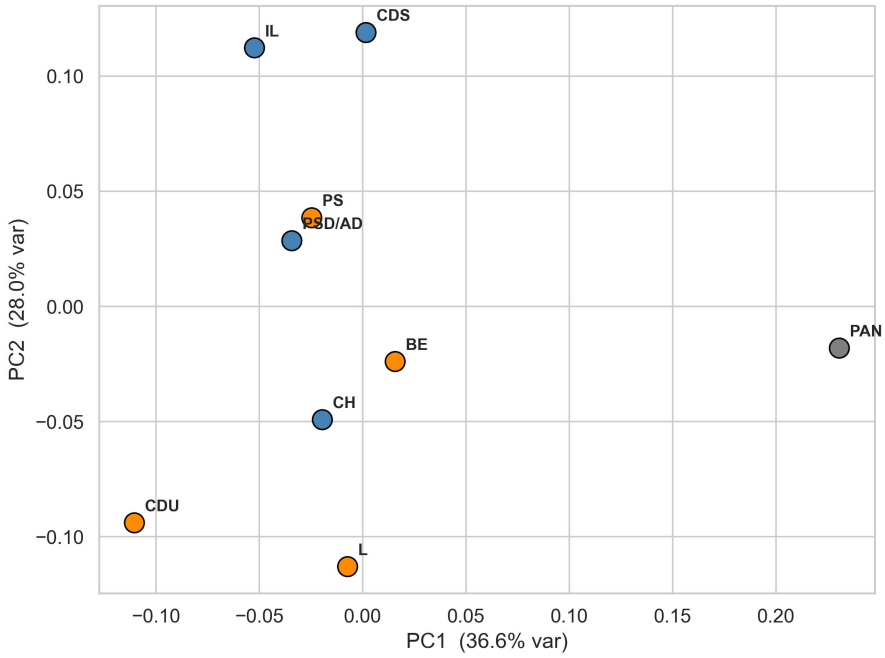
[21] On this topic, see Meguid (2005) or van Spanje and de Graaf (2018).

[22] See figure 15, in the appendix, for a representation of all elections on the same two dimensions, allowing for better comparison. Figure 14, on the other hand, maximizes the explanatory power of the used dimensions for each election.

Figure 12: Policy proposals by party (breakdown by debate opponent)

Notes: Total sum of all policy proposals made by each party's representatives in debates against each other party, across all elections.

Figure 13: Policy space (all elections combined)



Notes: Party proximity using a PCA(2) reduction of the 14-D policy space. Right-wing parties are presented in blue, left-wing parties in orange, and syncretic ones (PAN, in this case) in grey.

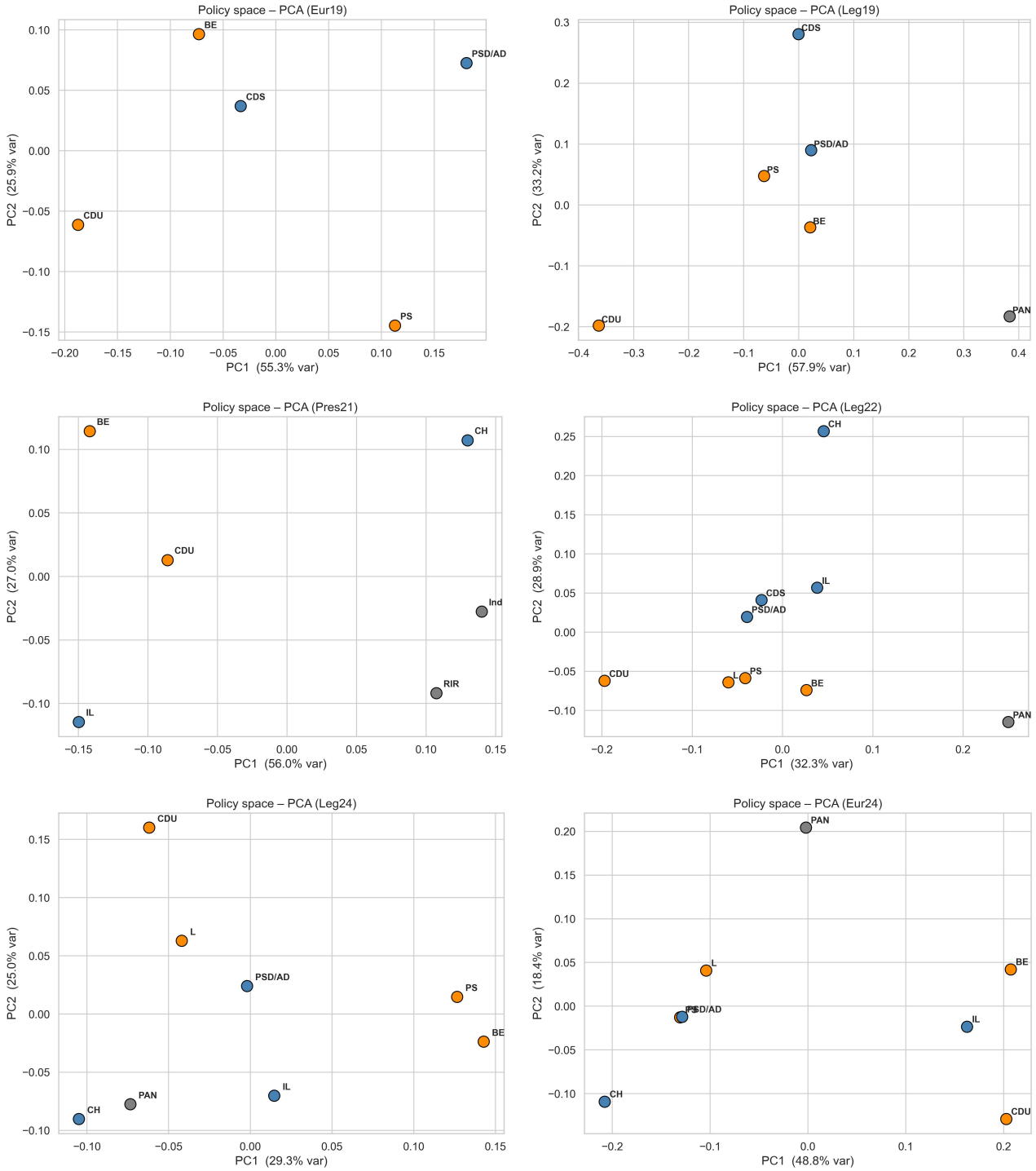
8.1 Ways forward

What is presented in this piece is as much a finished product, ready to be implemented, as it is a prototype to be expanded and improved upon. There are still many avenues for improving and documenting its performance: firstly, the ideal merge threshold for rare speakers when many speakers are present remains to be determined – an objective rule to determine this, perhaps using the number of speakers and total runtime of the media, would be a useful finding that could be implemented automatically (i.e. the number of speakers in each input is already autonomously determined by the pipeline) and could be determined with relatively brief testing.

Secondly, a structured evaluation of the performance impact of each new voice sample should be applied to the self-enrichment process, so as to determine the ideal maximum dimension of the voice bank per speaker, which might very well be above or below my chosen value of N=100. The centroid structure of voice profile aggregation should also be subject to a similar analysis, as my formulation is merely arbitrary and tailored specifically to the data at hand. The accuracy resulting from several possible formulations of the K rule should be compared, so as to determine the best progression structure (e.g. K=1 for N<10, K=2 for 11<N<20, etc.) and the ideal maximum value for K.

Additionally, a full audiovisual approach as in Kynych et al. (2024), using face-voice pairs, is likely to bring even better performance and should be tested (eventually in terms of computational demand vs. output accuracy). Finally, the use of LLMs can be as useful in supplementing the method as it is in interpreting its outputs – further integration could be employed to, for example, detect obvious mis-attributions and correspondingly, dynamically correct the voice bank.

Figure 14: Policy space (breakdown by election)



Notes: Party proximity using a PCA(2) reduction of the 14-D policy space. Dimensions are chosen individually per election. Right-wing parties are presented in blue, left-wing parties in orange, and syncretic ones (PAN, RIR and the independent candidates) in grey.

References

[1] Ajmera, J., Wooters, C. (2003). "A robust speaker clustering algorithm." 2003 IEEE Workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721). IEEE.

[2] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O. (2012). "Speaker Diarization: A Review of Recent Research." IEEE Transactions on Audio, Speech, and Language Processing 20(2): 356-370.

[3] Barberá, P., Rivero, G. (2015). "Understanding the Political Representativeness of Twitter Users." Social Science Computer Review 33(6): 712-729.

[4] Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., Mikhaylov, S. (2016). "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." American Political Science Review 110(2): 278-295.

[5] Bevan, S., John, P. (2016). "Policy representation by party leaders and followers: What drives UK Prime Minister's Questions?" Government and Opposition 51(1): 59–83.

[6] Bidwell, K., Casey, K., Glennerster, R. (2020). "Debates: Voting and Expenditure Responses to Political Communication." Journal of Political Economy 128 (8): 2880-2924.

[7] Ceron, T., Barić, A., Blessing, A., Haunss, S., Kuhn, J., Lapesa, G., Padó, S., Papay, S., Zauchner, P. F. (2024). "Automatic analysis of political debates and manifestos: successes and challenges." In: Cimiano, P., Frank, A., Kohlhase, M., Stein, B. (eds) Robust Argumentation Machines. RATIO 2024. Lecture Notes in Computer Science, vol 14638. Springer, Cham.

[8] Chaqués-Bonafont, L., Baumgartner, F. R. (2013). "Newspaper attention and policy activities in Spain." Journal of Public Policy 33(1): 65–88.

[9] Chung, J. S., Nagrani, A., Zisserman, A. (2018). "VoxCeleb2: Deep Speaker Recognition." arXiv:1806.05622.

[10] Clementson, D., Eveland, W. P., Jr. (2016). "When Politicians Dodge Questions: An Analysis of Presidential Press Conferences and Debates." Mass Communication and Society 19(4): 411-429.

[11] Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., Na, H. (2021). "ECAPA-TDNN Embeddings for Speaker Diarization." arXiv:2104.01466.

[12] Desplanques, B., Thienpondt, J., Demuynck, K. (2020). "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." arXiv:2005.07143.

[13] Di Leo, R., Zeng, C., Dinas, E., Tamtam, R. (2025). "Mapping (A)Ideology: A Taxonomy of European Parties Using Generative LLMs as Zero-Shot Learners." Political Analysis: 1-8.

[14] Druckman, J. N. (2003). "The Power of Television Images: The First Kennedy-Nixon Debate Revisited." Journal of Politics 65: 559-571.

[15] Ernst, N., Engesser, S., Büchel, F., Blassnig, S., Esser, F. (2017). "Extreme parties and populism: an analysis of Facebook and Twitter across six countries." Information, Communication & Society 20(9): 1347-1364.

[16] Fernandes, J. M., Debus, B., Bäck, H. (2021). "Unpacking the politics of legislative debates." European Journal of Political Research 60: 1032–1045.

[17] Fiscus, J. G., Ajot, J., Michel, M., Garofolo, J. S. (2006). "The rich transcription 2006 spring meeting recognition evaluation." In International Workshop on Machine Learning for Multimodal Interaction (pp. 309-322). Berlin, Heidelberg: Springer Berlin Heidelberg.

[18] Gish, H., Siu, M.-H., Rohlicek, R. (1991). "Segregation of Speakers for Speech Recognition and Speaker Identification." ICASSP 91: 873-876.

[19] Herbeck, D. A., Drury, S. A. M. (2022). "The first Kennedy-Nixon debate: the McKeesport Junto of 1947." Argumentation and Advocacy 58(3-4): 163-180.

[20] Kato, K., Purnomo, A., Cochrane, C., Saqur, R. (2024). "L(u)pin: Llm-based political ideology nowcasting." arXiv:2405.07320.

[21] Kynych, F., Cerva1, P., Zdansky, J., Svendsen, T., Salvi, G. (2024). "A lightweight approach to real-time speaker diarization: from audio toward audio-visual data streams." EURASIP Journal on Audio, Speech, and Music Processing 2024: 62.

[22] Laver, M., Benoit, K., Garry, J. (2003). "Extracting Policy Positions from Political Texts Using Words as Data." American Political Science Review 97(2): 311-331.

[23] McKinney, M. S., Carlin, D. B. (2004). "Political campaign debates". In Handbook of political communication research (pp. 203-234). Routledge.

[24] Meguid, B. M. (2005). "Competition Between Unequals: The Role of Mainstream Party Strategy in Niche Party Success." American Political Science Review 99(3): 347-359.

[25] Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., Besacier, L. (2006). "Step-by-step and integrated approaches in broadcast news speaker diarization." Computer Speech and Language 20: 303–330.

[26] Mingote, V., Ortega, A., Miguel, A., Lleida, E. (2024). "Audio-Visual Speaker Diarization: Current Databases, Approaches and Challenges." arXiv:2409.05659.

[27] Nagrani, A., Chung, J. S., Zisserman, A. (2017). "VoxCeleb: a large-scale speaker identification dataset." arXiv:1706.08612.

[28] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., Narayanan, S. (2022). "A review of speaker diarization: Recent advances with deep learning." Computer Speech and Language 72: 101317.

[29] Peeperkorn, M., Kouwenhoven, T., Brown, D., Jordanous, A. (2024). "Is Temperature the Creativity Parameter of Large Language Models?" arXiv:2405.00492.

[30] Self, J. W. (2005). "The First Debate over the Debates: How Kennedy and Nixon Negotiated the 1960 Presidential Debates." Presidential Studies Quarterly 35: 361-375.

[31] Silva, B. C., Proksch, S.-O. (2022). "Politicians unleashed? Political communication on Twitter and in parliament in Western Europe." Political Science Research and Methods 10: 776–792.

[32] Sharma, R., Narayanan, S. (2022). "Using Active Speaker Faces for Diarization in TV shows." arXiv:2203.15961.

[33] Siu, M-H., Yu, G., Gish, H. (1992). "An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers." IEEE International Conference on Acoustics, Speech, and Signal Processing 2: 189-192.

[34] Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. (2020). "The Dynamics of Political Incivility on Twitter." SAGE Open.

[35] Tranter, S. E., Reynolds, D. A. (2006). "An overview of automatic speaker diarization systems." IEEE Transactions on Audio, Speech, and Language Processing 14(5): 1557-1565.

[36] van Leeuwen, D. A., Konečný, M. (2007). "Progress in the AMIDA speaker diarization system for meeting data." International Evaluation Workshop on Rich Transcription. Berlin, Heidelberg: Springer Berlin Heidelberg.

[37] van Spanje, J., de Graaf, N. D. (2018). "How established parties reduce other parties' electoral support: the strategy of parroting the pariah." West European Politics 41(1): 1-27.

[38] Vijayasenan, D., Valente, F., Bourlard, H. "An information theoretic approach to speaker diarization of meeting data." IEEE Transactions on Audio, Speech, and Language Processing 17(7): 1382-1393.

[39] Wilkerson, J., Casas, A. (2017). "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." Annual Review of Political Science 20(1): 529-544.

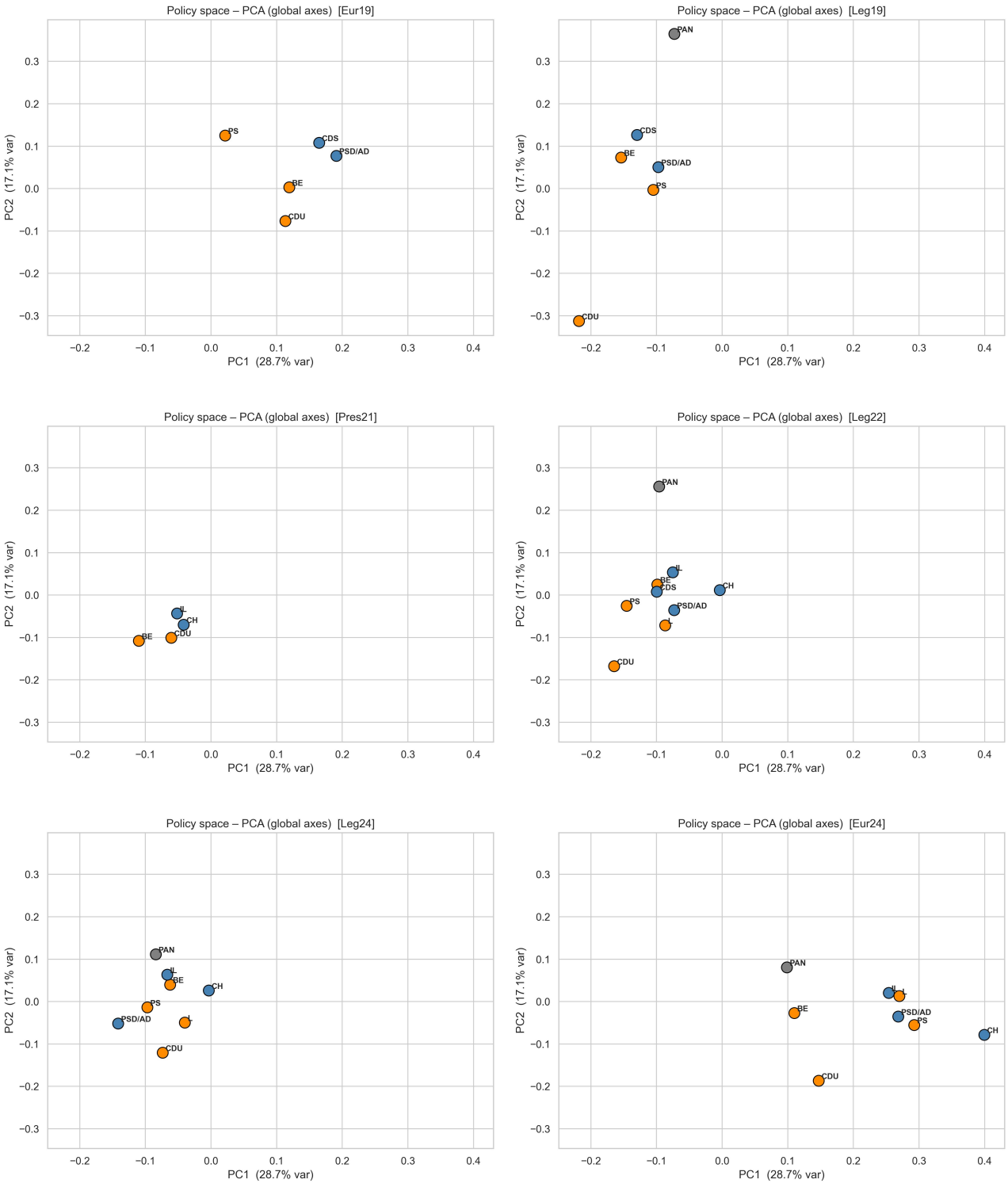
[40] Yehia, H., Rubin, P., Vatikiotis-Bateson, E. (1998). "Quantitative association of vocal-tract and facial behavior." Speech Communication, 26(1-2): 23-43.

[41] Zhu, X., Barras, C., Meignier, S., Gauvain, J.-L. (2005). "Combining speaker identification and BIC for speaker diarization." Interspeech'05: 2441-2444.

[42] Zulianello, M. (2014). "Analyzing party competition through the comparative manifesto data: some theoretical and methodological considerations." Quality & Quantity 48(3): 1723-1737.

Appendix

Figure 15: Policy space (breakdown by election) - fixed dimensions



Notes: Party proximity using a PCA(2) reduction of the 14-D policy space. Dimensions are chosen once for all elections, thus being the same across figures. Right-wing parties are presented in blue, left-wing parties in orange, and syncretic ones (PAN, RIR and the independent candidates) in grey.

A1: Video collection pipeline

```
import pandas as pd
from datetime import datetime
import os
import yt_dlp

#Get index - the spreadsheet illustrated in table 1
debates = pd.read_excel('debates.xlsx')

#Obtain an ID for the file names
debates['Title'] = debates['ID'] + " " + debates['P1'] + " x " + debates['P2']

#Drop missing links
debates = debates.dropna(subset=['Link'])

# Function to download both video and audio
def download_video_and_audio_yt_dlp(youtube_url, video_output_path, audio_output_path):
    # Download video (skip if exists)
    ydl_opts_video = {
        'format': 'bestvideo+bestaudio/best',
        'outtmpl': video_output_path,
        'ffmpeg_location': '/opt/anaconda3/envs/Faces/bin/ffmpeg',
    }

    # Download audio (skip if exists)
    ydl_opts_audio = {
        'format': 'bestaudio/best',
        'restrictfilenames': True,
        'outtmpl': audio_output_path + ".%(ext)s",  # yt-dlp will append .mp3 automatically
        'ffmpeg_location': '/opt/anaconda3/envs/Faces/bin/ffmpeg',
        'postprocessors': [{
            'key': 'FFmpegExtractAudio',
            'preferredcodec': 'wav',
            'preferredquality': '0',
        }],
    }

    # Skip if video is already present
    if not os.path.exists(video_output_path):
        print(f"Downloading video to {video_output_path}")
        with yt_dlp.YoutubeDL(ydl_opts_video) as ydl_video:
            ydl_video.download([youtube_url])
    else:
        print(f"Video already exists at {video_output_path}, skipping download.")

    # Skip if audio is already present
    if not os.path.exists(f"{audio_output_path}.wav"):
        print(f"Downloading audio to {audio_output_path}.wav")
        with yt_dlp.YoutubeDL(ydl_opts_audio) as ydl_audio:
            ydl_audio.download([youtube_url])
    else:
        print(f"Audio already exists at {audio_output_path}.wav, skipping download.")

# Iterate through the DataFrame and download both video and audio
for index, row in debates.iterrows():
    youtube_url = row['Link']
    video_title = row['Title']

    # Paths to save the video and audio (no .mp3 in audio file path)
    video_output_path = os.path.join('/Videos', f"{video_title}.mp4")
    audio_output_path = os.path.join('/Audio', f"{video_title}")  # yt-dlp will append .mp3 automatically

    # Ensure the directories exist
    os.makedirs('/Videos', exist_ok=True)
    os.makedirs('/Audio', exist_ok=True)

```

```
    # Download both video and audio
    download_video_and_audio_yt_dlp(youtube_url, video_output_path, audio_output_path)
```

A2: Segment embedding function and speaker embedding helper

```
def compute_spk_emb(wav, spkrec, sr=16000):
    """
    Returns a 192-D ECAPA embedding of the given WAV.
    If the snippet is shorter than MIN_EMB_SAMPLES or ECAPA fails,
    an all-zero vector is returned.
    """
    try:
        w, srate = torchaudio.load(wav)
    except Exception as err:
        print(f"[WARN] Couldn't load {wav}: {err}")
        return np.zeros(EMB_DIM, dtype=np.float32)

    if w.shape[0] > 1:                      # stereo to mono
        w = w.mean(0, keepdim=True)

    if w.shape[-1] < MIN_EMB_SAMPLES:       # too short for ECAPA
        return np.zeros(EMB_DIM, dtype=np.float32)

    if srate != sr:                         # resample if needed
        w = torchaudio.functional.resample(w, srate, sr)

    try:
        return spkrec.encode_batch(w.to(DEVICE_ECAPA)).squeeze().cpu().numpy()
    except Exception as err:
        print(f"[WARN] ECAPA failed on {wav}: {err}")
        return np.zeros(EMB_DIM, dtype=np.float32)

#####

def seg_emb(full_wav, spkrec, s, e):
    """
    Extracts an [s,e] snippet from full_wav, pads up to MIN_SNIP where possible,
    rejects it if it still is shorter than MIN_LEN_SAFE, and returns its
    speaker embedding (or None if rejected).
    """
    audio = AudioSegment.from_wav(full_wav)
    dur   = audio.duration_seconds

    # try to reach MIN_SNIP by symmetric padding
    if e - s < MIN_SNIP:
        pad = (MIN_SNIP - (e - s)) / 2
        s   = max(0, s - pad)
        e   = min(dur, e + pad)

    # still too short? skip
    if e - s < MIN_LEN_SAFE:
        return None

    # export snippet to tmp WAV for ECAPA
    with tempfile.NamedTemporaryFile(suffix=".wav", delete=False) as tf:
        tmp = tf.name
    audio[int(s * 1000) : int(e * 1000)].export(tmp, format="wav")

    emb = compute_spk_emb(tmp, spkrec)
    os.remove(tmp)
    return emb
```

A3: Voice sample embedding

```
def load_voice_refs(excel, voices_dir, spkrec):
    df=pd.read_excel(excel)
    vcols=[c for c in df.columns if c.lower().startswith("voice")]
    refs={}
    for _,row in df.iterrows():
        name=str(row.get("Nome",row.get("Name",f"Row{_}"))).strip()
        embs=[]
        for c in vcols:
            tag=row.get(c)
            if pd.notnull(tag):
                fp=Path(voices_dir)/f"{tag}.wav"
                if fp.exists():
                    embs.append(compute_spk_emb(fp,spkrec))
        refs[name]=np.mean(embs,0) if embs else None
    return refs
```

A4: Audio diarization function

```
def diarize_audio(segs, voice_refs, spkrec, wav, k_speakers, thr):
    embs=[]; diar=[]; known=set()
    for s in segs:
        e = seg_emb(wav, spkrec, s["start"], s["end"])
        if e is None: e = np.zeros(EMB_DIM, dtype=np.float32)
        embs.append(e)
    embs = np.stack(embs)

    if isinstance(k_speakers,int):
        lab,names,conf=cluster_and_name(embs,k_speakers,voice_refs)
        for i,s in enumerate(segs):
            lbl=names[lab[i]]
            if lbl.startswith("Unknown") or conf[i]<thr: lbl="Unknown"
            else: known.add(lbl)
            diar.append({**s,"speaker":lbl,"origin":"A","audio_conf":conf[i],"emb":embs[i]})
    else:
        for emb,s in zip(embs,segs):
            best,score=None,-1
            for n,ref in voice_refs.items():
                if ref is None: continue
                cs=cosine(emb,ref)
                if cs>score: best,score=n,cs
            lbl=best if best and score>=thr else "Unknown"
            if lbl!="Unknown": known.add(lbl)
            diar.append({**s,"speaker":lbl,"origin":"A","audio_conf":score,"emb":emb})
    return diar, list(known)
```

A5: Identify and merge aberrant speaker identifications

```
def merge_rare(diar, rare_pct=0.02, major_pct=0.05, thr=0.65):
    dur,embs,total=defaultdict(float),defaultdict(list),0
    for d in diar:
        span=d["end"]-d["start"]; total+=span
        dur[d["speaker"]]+=span; embs[d["speaker"]].append(d["emb"])
    rare ={s for s,t in dur.items() if t<rare_pct*total and s!="Unknown"}
    major={s for s,t in dur.items() if t>=major_pct*total and s!="Unknown"}
    if not rare or not major: return diar
    cent={s:np.mean(embs[s],0) for s in rare|major}
    rename={}
    for r in rare:
        best,score=None,-1
        for m in major:
            cs=cosine(cent[r],cent[m])
            if cs>score: best,score=m,cs
        rename[r]=best if score>=thr else "Unknown"
```

```
    for d in diar:
        if d["speaker"] in rename: d["speaker"]=rename[d["speaker"]]
    return diar
```

A6: Video fallback

```
def mouth_open(lm,ratio=0.2):
    if "top_lip" not in lm or "bottom_lip" not in lm: return False
    top,bot=np.array(lm["top_lip"]),np.array(lm["bottom_lip"])
    vert=np.linalg.norm(top.mean(0)-bot.mean(0))
    horiz=np.ptp(np.vstack((top,bot))[:,0])
    return horiz>1e-6 and vert/horiz>ratio

#####

def video_fallback(video, diar, photos, face_thr,
                   mouth_ratio=0.1, speak_thr=3,
                   min_frames=3, min_face_ratio=0.10):
    cap=cv2.VideoCapture(video); fps=cap.get(cv2.CAP_PROP_FPS) or 30
    W,H=int(cap.get(3)),int(cap.get(4))
    for seg in diar:
        if seg["speaker"]!="Unknown": continue
        stf,enf=int(seg["start"]*fps), int(seg["end"]*fps)
        cap.set(cv2.CAP_PROP_POS_FRAMES,stf)
        speak,counts=defaultdict(int),defaultdict(int)
        min_dist=1e9
        for _ in range(stf,enf+1):
            ok,frm=cap.read();
            if not ok: break
            rgb=cv2.cvtColor(frm,cv2.COLOR_BGR2RGB)
            locs=face_recognition.face_locations(rgb)
            encs=face_recognition.face_encodings(rgb,locs)
            lms =face_recognition.face_landmarks(rgb,locs)
            for loc,enc,lm in zip(locs,encs,lms):
                t,r,b,l=loc; fw,fh=r-l,b-t
                if fw<min_face_ratio*W or fh<min_face_ratio*H: continue
                best,dist=None,1e9
                for n,ref in photos.items():
                    d=np.linalg.norm(enc-ref)
                    if d<dist: best,dist=n,d
                if dist>=face_thr: continue
                if dist<min_dist: min_dist=dist
                if mouth_open(lm,mouth_ratio):
                    speak[best]+=1
                else: speak[best] = max(speak[best] - 1, 0)
                if speak[best]>=speak_thr:
                    counts[best]+=1
        if counts:
            winner,frames=max(counts.items(), key=lambda x:x[1])
            if frames>=min_frames:
                seg["speaker"]=winner
                seg["origin"]="V"
                seg["video_conf"]={"frames":frames,
                                   "ratio":frames/(enf-stf+1),
                                   "min_dist":min_dist}
    cap.release()

#####

def load_photo_refs(excel, img_dir):
    df=pd.read_excel(excel)
    cols=[c for c in df.columns if c.lower().startswith(("front","side","extra"))]
    refs={}
    for _,r in df.iterrows():
        name=str(r["Nome"]).strip(); vec=[]
        for c in cols:
```

```
            tag=r.get(c)
            if pd.notnull(tag):
                for ext in(".png",".jpg",".jpeg"):
                    fp=Path(img_dir)/f"{tag}{ext}"
                    if fp.exists():
                        fe=face_recognition.face_encodings(face_recognition.load_image_file(str(fp)))
                        if fe: vec.append(fe[0]); break
        if vec: refs[name]=np.mean(vec,0)
    return refs
```

A7: Self-learning dependencies

```
####Helpers

def _next_voice_column(df):
    vcols=[c for c in df.columns if c.lower().startswith("voice")]
    col=f"Voice{len(vcols)+1}"
    if col not in df.columns: df[col]=np.nan
    return col

######

def _next_seq_tag(base, voices_dir):
    nums=[]
    for p in Path(voices_dir).glob(f"{base}_*.wav"):
        suf=p.stem[len(base)+1:]
        if suf.isdigit(): nums.append(int(suf))
    n=max(nums)+1 if nums else 1
    return f"{base}_{n}"

#####

def _add_clip_to_faceindex(df,speaker,tag):
    name_col=_name_col(df)
    if speaker not in df[name_col].values:
        new_row={c:np.nan for c in df.columns}; new_row[name_col]=speaker
        df.loc[len(df)]=new_row
    row=df[name_col]==speaker
    vcols=[c for c in df.columns if c.lower().startswith("voice")]
    if df.loc[row,vcols].notnull().sum(axis=1).iloc[0]>=AUTO_REF_MAX_CLIPS:
        return
    for c in vcols:
        if pd.isnull(df.loc[row,c]).iloc[0]:
            df.loc[row,c]=tag; return
    df.loc[row,_next_voice_column(df)]=tag

#####

def _collect_candidate_segments(diar, min_conf, max_conf, min_dur):
    """
    Yield (speaker, start, end) for consecutive segments that
    meet *all* of these criteria:
       same named speaker
       duration >= min_dur
       min_conf <= mean audio_conf <= max_conf
    """
    i = 0
    while i < len(diar):
        seg = diar[i]; spk = seg["speaker"]
        if spk == "Unknown": i += 1; continue
        start, dur, confs = seg["start"], 0.0, []
        j = i
        while j < len(diar) and diar[j]["speaker"] == spk:
            dur += diar[j]["end"] - diar[j]["start"]
            confs.append(diar[j]["audio_conf"] or 0.0)
            j += 1
```

```
        mean_conf = np.mean(confs)
        if dur >= min_dur and min_conf <= mean_conf <= max_conf:
            yield spk, start, diar[j-1]["end"]
        i = j

#####Voicebank enrichment

def auto_enrich_voicebank(full_wav, diar, voices_dir, spkrec, face_df, self_enrich: bool = ENABLE_SELF_ENRICH):

    # ---------- global ON / OFF switch ----------------------------------------
    if not self_enrich:          # nothing will be written
        return
    """
    Harvest high-confidence segments and update the adaptive speaker models,
    but NEVER store more than AUTO_REF_MAX_CLIPS clips per speaker.
    """
    audio       = AudioSegment.from_wav(full_wav)
    voice_embs  = face_df.attrs["voice_embs"]        # {name: [emb, ]}
    voice_refs  = face_df.attrs["voice_refs"]        # {name: centroid(s)}

    for speaker, s, e in _collect_candidate_segments(
            diar,
            AUTO_REF_MIN_CONF,
            AUTO_REF_MAX_CONF,
            AUTO_REF_MIN_DUR):

        # ---------------- HARD QUOTA CHECK ----------------------------------
        if len(voice_embs.get(speaker, [])) >= AUTO_REF_MAX_CLIPS:
            # Nothing to do  we are at (or above) the quota for this speaker
            continue
        # --------------------------------------------------------------------

        # make an unused, sequential tag for this speaker
        base = re.sub(r"\W+", "_", speaker.lower()).strip("_")[:15]
        tag  = _next_seq_tag(base, voices_dir)

        # export snippet ------------------------------------------------------
        wav_path = Path(voices_dir) / f"{tag}.wav"
        audio[int(s * 1000): int(e * 1000)].export(str(wav_path), format="wav")

        # register the new clip in the face-index
        _add_clip_to_faceindex(face_df, speaker, tag)

        # embed & update the adaptive model -----------------------------------
        emb = compute_spk_emb(str(wav_path), spkrec)
        voice_embs.setdefault(speaker, []).append(emb)

        new_centroids          = _centroids_from_embs(voice_embs[speaker])
        voice_refs[speaker]    = new_centroids   # overwrite

        k = 1 if new_centroids.ndim == 1 else new_centroids.shape[0]
        print(f"{speaker:20s}    {len(voice_embs[speaker]):2d}/{AUTO_REF_MAX_CLIPS} "
              f"clips | {k} centroid(s)")

    # persist the updated dictionaries in the DataFrame
    face_df.attrs["voice_embs"] = voice_embs
    face_df.attrs["voice_refs"] = voice_refs
```

A8: GPT correction prompt

```
def build_prompt(context_segments: List[Dict], candidate_speakers: List[str]) -> str:
    candidate_lines = []
    for cand in candidate_speakers:
        info = FACE_INDEX.get(normalize(cand), {})
        role = info.get("role", "politician")
        partido = info.get("Partido", "unknown")
```

```
        candidate_lines.append(f"- {cand} ({role}, Partido='{partido}')")

    def fmt(seg, is_target=False):
        txt = seg["text"].strip().replace("\n", " ")
        return f'{seg["speaker"]}: {"###TARGET### " if is_target else ""}{txt}'

    excerpt = [
        fmt(s, is_target=(s["speaker"] == "Unknown")) for s in context_segments
    ]

    return textwrap.dedent(
        f"""
        You are an AI assistant helping to attribute debate turns.

        1. Focus on the line that contains ###TARGET###.
        2. Choose which candidate is speaking in that line.
        3. If <70 % confident, reply "Unknown".

        CANDIDATES:
        {os.linesep.join(candidate_lines)}

        CONTEXT:
        {os.linesep.join(excerpt)}

        Respond with a single JSON line:
        {{"speaker": "<name or Unknown>", "confidence": <0-1>, "reason": "<short>"}}"""
    ).strip()
```

A9: Policy proposal extraction

```
def extract_policy_proposals(json_path: Path, speaker: str) -> str | None:
    """
    Returns GPT answer (str) or None if <speaker> never speaks.
    The speaker matching is accent/whitespace/parenthesis tolerant.
    """
    segments = json.loads(json_path.read_text(encoding="utf-8"))

    speaker_text = " ".join(
        seg["text"]
        for seg in segments
        if _norm_name(seg.get("speaker", "")) == _norm_name(speaker)
    )
    if not speaker_text.strip():
        return None  # speaker not present after normalisation

    # ---- build messages ----
    msgs = [{
        "role": "system",
        "content": (
            "You are an assistant that extracts POLICY PROPOSALS from a "
            f"Portuguese political debate. Work ONLY with the provided "
            f"utterances by {speaker}. Do NOT invent facts. "
            "Output exhaustive bullet points; for each proposal give "
            "a <=20-word paraphrase plus the exact Portuguese quote(s) "
            "as evidence."
        ),
    }]

    for chunk in split_by_tokens(speaker_text):
        msgs.append({"role": "user",
                     "content": f"[{speaker.upper()} UTTERANCES]\n{chunk}"})

    msgs.append({"role": "user",
                 "content": "List every concrete policy proposal made by the speaker in this debate."})

    resp = client.chat.completions.create(
```

```
        model=MODEL_NAME,
        messages=msgs,
        temperature=0.0,
    )
    return resp.choices[0].message.content.strip()
```

A10: Policy proposal classification

```
# 4) GPT helper  deduplicate + classify proposals from one file
def classify_proposals(raw_txt: str,
                       categories: list[str]) -> list[dict[str, str]]:
    """
    Returns a list of dicts: {proposal, category}
    Category must be chosen from <categories>.
    The function tolerates either a JSON array or an object {"data": [...]}
    coming back from the model.
    """
    sys_msg = (
        "You are an assistant that classifies policy proposals extracted "
        "from Portuguese political debates.\n"
        "Task:\n"
        "1. Deduplicate near-identical proposals (different wording, same idea: one entry).\n"
        "2. Assign ONE best-fit category (from the allowed list) to each distinct proposal.\n"
        "3. Output valid JSON ONLY, either\n"
        "   a) a plain JSON array, or\n"
        "   b) an object with a single key 'data' whose value is that array.\n"
        "Each array element: {\"proposal\": \"...\", \"category\": \"...\"}\n\n"
        "Allowed categories:\n- " + "\n- ".join(categories)
    )

    msgs = [
        {"role": "system", "content": sys_msg},
        {"role": "user",   "content": raw_txt},
        {"role": "user",   "content": "Return the JSON now."}
    ]

    resp = client.chat.completions.create(
        model="gpt-4.1",
        messages=msgs,
        temperature=0.0,
    )

    content = resp.choices[0].message.content.strip()
```