Identity Emergence in the Context of Vaccine Criticism in France

Melody Sepahpour-Fard[1], Michael Quayle[1,2], Padraig MacCarron[1], Shane Mannion[1], Dong Nguyen[3]

[1]University of Limerick, Ireland
[2]University of KwaZulu-Natal, South Africa
[3]Utrecht University, Netherlands
melody.sepahpourfard@ul.ie, mike.quayle@ul.ie, padraig.maccarron@ul.ie, shane.mannion@ul.ie, d.p.nguyen@uu.nl

Abstract

This study investigates the emergence of collective identity among individuals critical of vaccination policies in France during the COVID-19 pandemic. As concerns grew over mandated health measures, a loose collective formed on Twitter to assert autonomy over vaccination decisions. Using analyses of pronoun usage, outgroup labeling, and tweet similarity, we examine how this identity emerged. A turning point occurred following President Macron's announcement of mandatory vaccination for health workers and the health pass, sparking substantial changes in linguistic patterns. We observed a shift from first-person singular (*I*) to first-person plural (*we*) pronouns, alongside an increased focus on vaccinated individuals as a central outgroup, in addition to the media and President Macron. This shift in language patterns was further reflected in the behavior of new users. An analysis of incoming users revealed that a core group of frequent posters played a crucial role in fostering cohesion and shaping norms. New users who joined during the week of Macron's announcement and continued posting afterward showed an increased similarity with the language of the core group, contributing to the crystallization of the emerging collective identity. By leveraging large-scale social media data and computational methods, we provide insights into the mechanisms through which resistance movements solidify their identity online in response to policy changes.

Introduction

Group identity, i.e., the part of our identity tied to group membership, is a fundamental aspect of human experience that shapes our emotions and behaviors (Baumeister and Leary 1995). It facilitates the motivation and collective power to act as a group (Baumeister and Leary 1995; Brewer 1991), making it central to collective action and social movements (Reicher 1996; Simon et al. 1998). These socially negotiated collective identities influence and reshape our societies. Therefore, understanding the emergence of collective identities is essential to inform policy makers and activists about the mechanisms driving these identities and their potential societal impacts.

Social media platforms like Twitter (now X) provide a unique window into the dynamics of groups and social movements, offering vast amounts of naturalistic data that capture real-time interactions among users. Previous research has leveraged social media to explore a variety of social phenomena, including polarization (Tucker et al. 2018), social influence (González-Padilla and Tortolero-Blanco 2020), and self-presentation (Sepahpour-Fard and Quayle 2022; Sepahpour-Fard et al. 2023). Furthermore, social media enables researchers to observe how individuals express their identities, form ingroups and outgroups, and respond to significant events such as the COVID-19 pandemic and related policy changes.

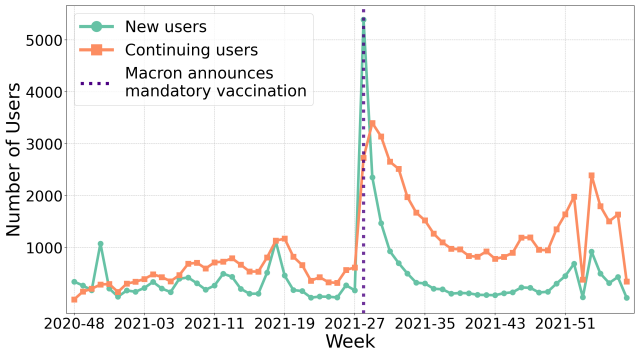


Figure 1: Number of new and continuing vaccine-critical users by week, where a new user is posting for the first time and a continuing user has posted at least once before. Macron's announcement on mandatory vaccination for health workers and the health pass coincided with a surge of new users joining the conversation. This paper investigates how his speech triggered the emergence of a collective identity within vaccine-critical communities.

While numerous studies have focused on the impact of COVID-19, particularly in terms of social media influence (González-Padilla and Tortolero-Blanco 2020) and misinformation spread (Hossain et al. 2020), relatively few have examined the anti-vaccine identity (Motta et al. 2023; Kadić-Maglajlić, Lages, and Pantano 2024). Moreover, to our knowledge, none have specifically focused on the *process* of collective-identity emergence among vaccine-critical individuals.

France's high level of vaccine skepticism makes it a particularly relevant context for studying the emergence of

vaccine-critical identity. Identified by the Wellcome Global Health Monitor as one of the most vaccine-skeptical countries in the world (Gallup 2019), this skepticism is rooted in historical distrust of the pharmaceutical industry, past vaccine controversies, and a strong cultural emphasis on individual freedom (Ward et al. 2022). The introduction of the health pass and mandatory vaccination for health workers by President Emmanuel Macron on July 12, 2021, intensified these sentiments, with many viewing it as an overreach of government authority and an infringement of personal freedom (Faccin et al. 2022).

COVID-19 vaccine policies (e.g., mandates and passes) likely played a key role in strengthening the collective identity of vaccine-critical individuals by framing opposition as resistance to coercive government measures. Recent research has highlighted the consequences of these policies. In France, even among vaccinated individuals, doubts about the vaccine increased from 44% to 61% following the health pass implementation (Ward et al. 2022). Furthermore, these policies may have exacerbated public distrust in scientific institutions and policymakers, reinforced social polarization, and fueled resistance to vaccination efforts (Bardosh et al. 2022). These sentiments have fueled widespread protests (Bronner and Mandard 2021) and led to the formation of online communities dedicated to opposing vaccine mandates and promoting skepticism (Peretti-Watel et al. 2020). Only a few studies have analyzed the content of French anti-vaccine tweets and the users they attract (Faccin et al. 2022), the impact of mandatory vaccination on vaccination rates, and anti-vaccine arguments (Sauvayre 2023; Gable, Sauvayre, and Chauvière 2023). This gap in the literature on the study of identity emergence may stem from the challenges associated with studying such a fluid and complex phenomenon (Morselli et al. 2023). Computational methods and large-scale social media data offer a valuable opportunity to explore the dynamic emergence of identity and its underlying mechanisms.

Present Work. We explore the emergence and evolution of collective identity among Twitter users critical of COVID-19 vaccines and related policies in France, highlighting the role of events catalyzing shifts in group identity and the importance of language in these processes.

Data and Methods. We analyze a dataset of 338,641 tweets posted between December 2020 and January 2022 that contains specific hashtags associated with vaccine criticism in France. First, we examine pronoun usage by measuring the prevalence of first-person *singular* (e.g. *je*/*I*) and *plural* (e.g., *nous*/*we*) pronouns. Then, we identify outgroup labels based on their cosine similarity to third-person plural pronouns (i.e., *they* and *them*) in word embedding models (Mikolov et al. 2013). Finally, we use cosine similarity of sentence embeddings (Martin et al. 2020) and the Fightin' Words method (Monroe, Colaresi, and Quinn 2008) to investigate the language of different groups of users.[1]

Results. Our findings revealed that President Macron's speech, on July 12th, 2021, acted as a catalyst, leading to the consolidation of collective identity. Figure 1 shows the surge in the number of users criticizing COVID-19 vaccines and related policies following his speech. We identified several key components of collective identity emergence:

(1) Pronoun use: Before Macron's announcement, tweets predominantly used first-person singular pronouns, reflecting individual perspectives. Following the announcement, there was a marked shift towards first-person plural pronouns, indicating an emerging sense of collective identity.

(2) Outgroups: Our analysis showed that users initially targeted political authorities in general (e.g., the government), but following Macron's speech the focus shifted toward vaccinated individuals, media actors, and Macron himself as salient outgroups.

(3) Incoming members: Users who joined the discussion during the week of Macron's speech—when the influx of new users was highest (see Figure 1)—and continued posting afterward exhibited increased similarity with established users. This trend may indicate a tendency to conform to the group's linguistic norms and assert their identity within the movement (Nguyen and P. Rosé 2011).

Implications. Studying the emergence of collective identity among vaccine-critical individuals has important implications. This group actively protested vaccination policies, and understanding how mandatory measures may have strengthened opposition can help policymakers design better strategies to manage resistance and improve engagement with controversial policies. Broader implications include the role of online platforms in shaping group identities and social movements, making this study relevant to both academic research and policy-making.

Related Work

We contextualize our analysis by reviewing key studies and theories on collective identity. First, we define collective identity, then examine the shift from personal to collective identity and the role of outgroups. Finally, we discuss research on the dynamics of joining established groups.

Collective Identity: Definition

Definitions of collective identity vary across the literature. Polletta and Jasper (2001) define it as "the individual's cognitive, moral, and emotional connection with a broader community, category, practice, or institution" (p.285).[2] Other scholars, such as Snow (2001), view collective identity as an interindividual process, whereby it emerges through interactions and actions (Snow 2001), as well as through shared interests, experiences, and solidarity, leading to a shared definition (Taylor and Whittier 1992) and a sense of belonging to a group (Smithey 2009). The identification process can lead to self-stereotyping, in which individuals amplify their similarities with the ingroup, adopt prototypical behaviors, and conform to the ingroup norms (Simon and Hamilton 1994;

[1] The code for these analyses can be found on GitHub: melodys-pr/identity_emergence_vaccine

[2] This definition, with its focus on the individual's perspective, also aligns with the concept of social identity as defined by social psychologists (Tajfel and Turner 1979). In this study, we draw from both the Sociology and Social Psychology literature to explore group identity processes.

Tajfel and Turner 1979; Moreland 1985), e.g., through common language (Flesher Fominaya 2010).

In relation to social movements, collective identity helps conceptualize how individuals unite, coordinate, and commit within a movement, as well as how these movements emerge and persist (Flesher Fominaya 2010). It is regarded as a crucial element for the cohesion and success of social movements (Melucci 1980, 1985, 1988). Previous research has explored collective identity on social media in various movements, including the Yellow Vests movement in France (Lüders, Dinkelberg, and Quayle 2022; Morselli et al. 2023), the Iranian 'My Stealthy Freedom' movement (Khazraee and Novak 2018), the responses to Russia's invasion of Ukraine (O'Reilly et al. 2024), and the Stop the Steal campaign following the 2020 US presidential election (Spann et al. 2023). These studies have employed both qualitative methods (Khazraee and Novak 2018) and quantitative approaches such as topic modeling (Morselli et al. 2023; Spann et al. 2023) and network analysis (Spann et al. 2023). For example, Spann et al. (2023) analyzed the Stop the Steal campaign on Twitter by using topic modeling to track the evolution and alignment of discussion topics, and social network analysis to evaluate the structural cohesion of the community over time. Additionally, despite some exceptions (Rousseau and Van Der Veen 2005; Drury and Reicher 2000), empirical studies, particularly in Social Psychology, have often treated collective identity as a predefined construct guiding collective action (Van Zomeren, Postmes, and Spears 2012; van Zomeren, Kutlaca, and Turner-Zwinkels 2018). In contrast, the present study uses multiple linguistic indicators—such as pronoun use, outgroup labeling, tweet similarity, and distinctive words—alongside large-scale social media data to capture the dynamic and evolving nature of collective identity through language.

From a Personal to a Collective Identity

The emergence of collective identity marks a shift from a personal (i.e., "I") to a collective (i.e., "we") locus of self-definition (Brewer and Gardner 1996; Taylor and Dube 1986). The choice between first-person singular and plural pronouns reflects individuals' relationship with their audience (Maitland and Wilson 1987), showing how language can be a central part of identity formation in interaction (Labov 2011). The first-person plural pronoun "we", in particular, conveys a sense of inclusion and belonging, activating and emphasizing shared identity among speakers (Pennebaker 2011; Brewer and Gardner 1996).

Several studies have examined the relationship between first-person plural pronouns and group-identity orientations (Pennebaker and Chung 2014; Lee et al. 2020). For instance, Lee et al. (2020) examined group identity within a K-pop fandom on Facebook and observed that the increased use of "we" over "I" for self-referencing not only predicted higher levels of group interaction, but also facilitated the consolidation of a cohesive group identity by fostering a shared sense of belonging and collective action within the fandom. Similarly, Pera and Aiello (2024) operationalize collective identity in YouTube discourse on veganism by deriving a Collective Identity Index based on the relative use of first-person singular and plural pronouns, further underscoring the central role of pronouns in capturing shifts from personal to collective identity in online discourse.

Outgroups and Comparison

When defining one's identity, two criteria of comparison are involved: sameness and distinctiveness (Pickett and Leonardelli 2006). While sameness can be achieved within the ingroup, distinctiveness is achieved through comparison with relevant outgroups (Brewer 1991). When in conflict with an outgroup, the perceived opposition makes groups tend towards ideas and actions in reaction to each other (Van Stekelenburg 2013). For instance, Drury and Reicher (2000) found that confrontations with an outgroup, such as the police during protests, could unite otherwise divided protest participants and radicalize them. Research has also shown how institutionalized rejection of a group based on identity can spark major social movements (Marx 1998).

Relative deprivation, the perception of being unfairly disadvantaged compared to others, is another factor influencing social movements, intergroup attitudes, and collective action (Lüders et al. 2021; Guimond and Dambrun 2002). The theory of relative deprivation states that people do not make absolute judgment on fairness but rather perceive a situation as fair or unfair in comparison to outgroups (Crosby 1976). This comparison can be vertical, with a group having a position of power (e.g., the "elite" or an authority), or horizontal, with groups of comparable status (e.g., vaccinated people for non-vaccinated ones; Lüders et al. 2021). Research has shown relative group deprivation, both vertical and horizontal, to be positively related to protest participation (Lüders et al. 2021).

Established and Incoming Users

In a social movement, incoming members may adopt the behavior of established members to enhance their sense of belonging to the group. When individuals enter a group, they are met with certain expectations that may influence their attitudes and behaviors, regulating their assimilation to the group (Moreland 1985). Established ingroup members will define the norms and therefore how to think and act for incoming group members (Livingstone et al. 2011).

Additionally, incoming members' motivation to join the group can vary. High identifiers, who are strongly motivated to be part of the group, are more likely to exert extra effort to be accepted as legitimate members (Branscombe et al. 1999). They will often engage in prototypical behaviors and align their language with ingroup norms to gain favorable recognition (Noel, Wann, and Branscombe 1995), while using derogatory language to differentiate themselves from the outgroup (Branscombe et al. 1999). In contrast, low identifiers, who are less invested in the group, may exhibit less alignment (Branscombe et al. 1999). For instance, Danescu-Niculescu-Mizil et al. (2013) investigated how incoming users integrate into communities with established norms and found that linguistic conformity with ingroup norms predicts incoming users' commitment to the group.

Dataset

To explore the emergence of collective identity, we focus on individuals criticizing COVID-19 vaccines and related policies in France during the pandemic. In this section, we describe our Twitter data collection and data preparation steps. Although we report this analysis and give examples in English, raw data and the analysis pipeline were in French.

Data Collection

Based on previous research that widely uses hashtags for Twitter data collection (Reavley and Pilkington 2014; Yardi and Boyd 2010; Linabary, Corple, and Cooky 2020), and given that hashtags serve as rallying points in social movements and enable a more precise targeting of specific topics and audiences (Conover et al. 2021), we choose to base our data collection on hashtags. We compile a list of hashtags associated with vaccine criticism through a manual snowball search on Twitter (i.e., by searching known hashtags and gathering co-occurring hashtags from tweets seemingly posted by vaccine-critical users), as well as from various websites and blog posts.[3] The finalized set contains the following hashtags: #PassSanitaire (health pass), #VaccinationObligatoire (mandatory vaccination), #antivax, #antivaccin, #antivaxx, #NonAuVaccin (no to vaccine), #JeNeMeFaisPasVacciner (I am not getting vaccinated), #JeNeMeFeraisPasVacciner (I won't get vaccinated), #PassDeLaHonte (shame pass), #NonAuPassDeLaHonte (no to the shame pass), #NonAuPassSanitaire (no to the health pass), #NonAuPassVaccinal (no to the vaccination pass), #DictatureSanitaire (health dictatorship), #StopDictatureSanitaire (stop health dictatorship), #NousSommesDesMillions (we are millions).

Using the Twitter API v2 for academic research, we collect all tweets and retweets containing at least one of the hashtags from December 2020 to January 2022. Our dataset encompasses a total of 2,177,538 posts. In our analysis, we specifically focus on original tweets and exclude retweets (i.e., posts starting with "RT"), because we want to analyse the active participation in the creation of linguistic content. This results in a dataset of 402,836 tweets authored by 38,998 unique users.

Data Preparation

We clean the data by eliminating mentions, URLs, duplicated punctuation, extra spaces, and tweets consisting solely of punctuation, and we remove duplicates. Additionally, we remove tweets censored by Twitter for Terms of Service violation since their content was deleted and replaced by the text "[User] account is temporarily unavailable because it violates the Twitter Media Policy. Learn more.", resulting in a refined dataset comprising 401,067 tweets authored by 38,836 unique users.

Following a manual review of the dataset, we notice that the collected dataset contains tweets from pro-vaccination users who promote vaccination and criticize vaccine-hesitant users. As we study the emergence of collective identity within the group formed by people criticizing COVID-19 vaccines and related policies, we need to remove tweets posted by other groups from our dataset. We identify tweets that do not align with our study focus, including: a) pro-vaccination tweets explicitly advocating for the vaccine or criticizing vaccine hesitancy; b) tweets seemingly not representing individual users; c) tweets criticizing both pro- and anti-vaccine positions; d) unrelated tweets, i.e., tweets using at least one of the hashtag used for data collection but with a content unrelated to COVID-19 or vaccination. To remove those tweets, we train a classifier (Martin et al. 2020) on manually classified tweets (1% of the total dataset). The classifier achieved a precision of .97, a recall of .96, and an F1-score of .97 for predicting the class of interest: tweets related to the criticism of COVID-19 vaccines and related policies. We use our classifier on the remaining unlabeled dataset and filter out non-relevant tweets (15.56% of tweets). Our final dataset for all subsequent analyses consists of 338,641 tweets and 27,016 users (see Appendix for more details on classification).

To evaluate the influence of automated accounts, we run the Botometer (Davis et al. 2016; Sayyadiharikandeh et al. 2020) on the 27,016 unique vaccine-critical accounts. The Botometer provides a continuous score (0–1) indicating bot-likeliness, with 0.5 frequently used as a threshold in prior work (Badawy, Ferrara, and Lerman 2018; Varol et al. 2017; Shao et al. 2018). We obtained scores for 21,872 accounts (81%), with an average score of 0.21 (median = 0.16). Only 6.07% of accounts exceeded a bot score of 0.5, and these accounts have posted only 2.35% of tweets. To evaluate the impact of these accounts on our findings, we re-ran our main analyses—pronoun use, outgroup label prevalence, cosine similarity between groups of users, and outgroup labels by author group—after excluding accounts with bot scores > 0.5. The results were unchanged (see Appendix for figures and detailed results), suggesting that automated activity has negligible impact on our findings.

From *I* to *We*

Pronouns, particularly first-person singular and plural pronouns, reveals the authors' relationships with their audience. We examine the prevalence of these pronouns in our data and how they reflect the emergence of collective identity.

Methods

We compile a comprehensive list of pronouns categorized by the person to whom they refer: first-person singular (1SG,

[3]The blogpost "Les mouvements anti pass et antivax sur les réseaux sociaux en France: les hashtags" ("The anti-pass and antivax movements on Social Media in France: the hashtags") written by Christophe Asselin's and posted on January 25, 2022, provided a set of hashtags used by French antivax and anti-pass users that we included in our list. Internet archive: http://web.archive.org/web/20230209102632/https://blog.digimind.com/fr/agences/mouvements-anti-pass-antivax-reseaux-sociaux-france-etude-hashtags

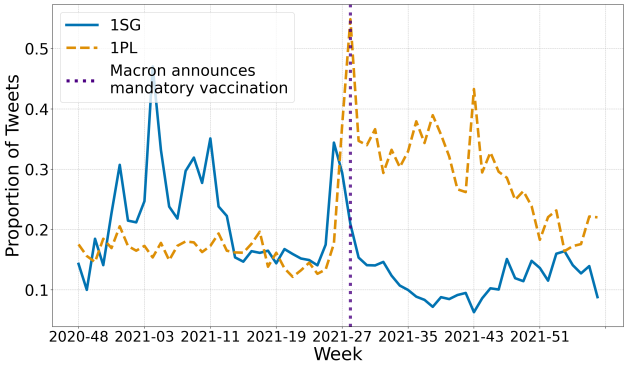


Figure 2: Pronoun use across weeks, considering two groups of pronouns: first-person singular (1SG) and first-person plural (1PL) pronouns. Before Macron's speech, users mostly used 1SG pronouns while they used more 1PL pronouns after.

e.g., *je*/*I*)[4] and first-person plural (1PL, e.g., *nous*/*we*).[5] Subject (e.g., *I*), object (e.g., *me*), and possessive (e.g., *mine*) pronouns are included in our analysis. To analyze the trajectory of pronoun usage, we identify occurrences of pronoun groups based on two criteria:

- if any pronoun (subject, object, or possessive) appears within the body of a tweet, excluding hashtags;
- if a subject pronoun (i.e., *I* or *we*) is at the beginning of a hashtag.

If either of these conditions is met, the tweet is labeled as 1 (present) for that pronoun group; otherwise, it is labeled as 0 (absent). A single tweet can contain both pronoun groups. We then use proportion z-tests to compare pronoun use before and after Macron's July 12, 2021, announcement on mandatory vaccination and the health pass.

Results

Figure 2 shows the evolution of the use of pronouns across weeks. Initially, users mostly used first-person singular pronouns such as "I" or "my". The major shift in pronoun use happened the week before Macron's speech (2021-27), with a peak on the week of the speech (2021-28). First-person plural pronouns such as "we" or "us" became the most used pronouns and they were constantly the most used pronoun group until the end of the time frame.

Results show a important shift in pronoun use before and after Macron's speech. Before the speech, 1SG pronouns were used in 22% of tweets, dropping to 13% afterward, while 1PL pronouns increased from 19% to 31% (Figure 12, Appendix). These changes were statistically significant, with 1SG pronouns showing a decrease (Z=60.45, p=0.0) and 1PL pronouns showing an increase (Z=-69.49, p=0.0).

[4]Full list of first-person singular pronouns: *je* (I), *j'* (I), *me* (me), *moi* (me), *mien* (mine), *mienne* (mine), *miens* (mine), *miennes* (mine), *ma* (my), *mon* (my)

[5]Full list of first-person plural pronouns: *nous* (we/us), *notre* (our), *nos* (our), *nôtre* (ours), *nôtres* (ours)

To assess the robustness of the observed pronoun shift and the influence of hashtags, we conducted a series of supplementary analyses. First, we removed hashtags from tweets entirely. In this case, the shift from 1SG to 1PL pronouns was not maintained (see Appendix, Figure 6A), indicating that hashtags play a central role in pronoun prevalence. Second, we examined hashtags in isolation (i.e., removing non-hashtag tokens). In this case, the shift from 1SG to 1PL pronouns was maintained (see Appendix, Figure 6B). Finally, we excluded only the hashtags used for data collection while retaining all other hashtags. Here, the results persisted: we continued to observe a clear shift from 1SG to 1PL pronouns (see Appendix, Figure 6C). Overall, these robustness checks indicate that while hashtags shaped the pronoun dynamics, the observed shift cannot be explained only by artifacts of the data collection process. Additional details and supplementary figures are provided in the Appendix.

The importance of 1PL ("we") pronouns in constructing collective identity is well-documented (Maitland and Wilson 1987; Pennebaker 2011; Lee et al. 2020; Papapavlou and Sophocleous 2009; Smith, Gavin, and Sharp 2015). The shift from 1SG to 1PL pronouns in our findings suggests the emergence of a collective identity among French-speaking Twitter users critical of COVID-19 vaccines and related policies. Before Macron's speech, users primarily expressed individual perspectives through 1SG pronouns. However, the July 12, 2021, announcement appears to have triggered a shift towards defining oneself in relation to the ingroup (Turner et al. 1994), as reflected in the increased use of 1PL pronouns.

Who Are *They*?

In this section, we identify outgroup references and examine how outgroups shift over time, particularly in response to President Macron's July 12, 2021, speech.

Methods

To analyze how users framed their outgroups, we implemented a pipeline using word embeddings validated by a human coder and a large language model (GPT-5). Word embeddings can capture distributional similarity at the word level, allowing us to identify clusters of terms functioning as outgroup markers. Similar approaches have been applied in prior work on identity labeling, where seed identity terms were expanded through embeddings to detect semantically related vocabulary (Yoder et al. 2023). Additionally, recent work also shows the highest quality labels emerge when GPT and human annotations are combined (He et al. 2024).

Tweets were preprocessed by removing nonalphabetic characters (except hashtags) and converting text to lowercase. Because we observed frequent occurrences of the phrase "non vacciné" (non-vaccinated) and synonyms wherein "non" precedes "vacciné" to negate it, we treat this phrase and similar ones as a distinct token by inserting an underscore between the two words (i.e., "non_vacciné"). We train two distinct word embedding models using Word2Vec (Mikolov et al. 2013): one for tweets posted before July 12, 2021, and another for those posted from

July 12, 2021, onwards (vector_size=100, window=5, and min_count=5). This division is motivated by the sharp increase in tweet volume after this date (see Figure 1).

For each embedding model, we retrieved the 50 most similar words (by cosine similarity) to the pronouns "ils" (they) and "eux" (them), restricting to words occurring at least 20 times. This was done separately for pre- and post-July 12 tweets. We then went through these words manually to collect those potentially referring to a group of people (i.e., candidate outgroup label) and, in parallel, prompted GPT-5 to perform the same task (Prompt: "Which of these words could refer to an outgroup (i.e., people distinct from the group of vaccine-critical individuals)? Give me a Python list.") The sets of terms identified by both methods were then combined, resulting in 38 candidate outgroup labels.

Next, we expanded the candidate set by examining the 50 nearest neighbors of the words identified in the first round. Again, both manual inspection and GPT-5 were applied to select words that could potentially serve as outgroup labels, using the same prompt as above. In addition, a small number of contextually salient terms observed during close reading of tweets were added ("talibans", "ayatollahs", "vacc"/vax, "hypnotisés"/hypnotized, "vaccinaux"/vaccine-related, "endoctrinés"/indoctrinated, "vaxxés"/vaxxed, "voyous"/thugs, and "terroristes sanitaires"/health terrorists), along with the singular forms of présidents (presidents) and gouvernements (governments). We then merged these lists of terms, resulting in a final candidate set of 231 words.

For each candidate outgroup label, 20 random tweets containing the term were randomly sampled. Both manual coding and GPT-5 were used to assess whether the term referred to an outgroup. GPT-5 was prompted with the following instruction: "These tweets are written by people criticizing vaccine-related policies or resisting vaccine mandates or non-vaccinated people. In how many of these tweets do they use the term <WORD> to refer to people in France who are distinct from those criticizing vaccine-related policies or resisting vaccine mandates or non-vaccinated people? Return an explanation for each tweet and the final count." Agreement between human and GPT-5 coding was moderate (Fleiss' Kappa = 0.49; Landis and Koch 1977). The moderate agreement may be due to the frequent use of irony and the brevity of tweets (Sauvayre, Vernier, and Chauvière 2022). Their counts were also moderately to strongly and significantly correlated (Pearson r = 0.65, Spearman ρ = 0.62, both $p < 0.001$).

Finally, only labels where both methods agreed that at least 15 of 20 sampled tweets ($\geq$75%) referred to an outgroup were retained. Sensitivity analyses on this threshold, reported in the Appendix, showed that varying the cutoff did not substantially affect the results.

To assess whether the prevalence of specific outgroup terms shifted following President Macron's speech, we conducted two-sample z-tests for proportions. For each outgroup label, we compared its relative frequency in tweets posted before July 12, 2021, with its frequency afterward.

Results

We identified a total of 94 outgroup labels through the combined word embedding and validation pipeline (see Appendix for the full list of outgroup labels). To assess how their prevalence changed following President Macron's July 12, 2021, speech, we compared their normalized frequencies in tweets posted before and after this date. Using two-sided two-sample z-tests for proportions, we found that 52 labels exhibited statistically significant changes in prevalence.

Figure 3 visualizes the difference in normalized frequencies (after–before) for these 52 labels. The most striking finding is that vaccinated people emerged as a central outgroup following Macron's announcement. The label "vaccinés" (vaccinated) showed the strongest increase in prevalence, in addition to the labels "spikés" (spiked), "dosés" (dosed), and "vaxxinés" (vaccinated). It suggests that opposition crystallized around distinguishing between vaccinated and non-vaccinated individuals.

In addition, the media became a salient outgroup after the speech. Terms such as "merdias" (a derogatory portmanteau meaning "shitty media"), "journalistes" (journalists), and "journaleux" (pejorative term for journalists) all increased significantly. This pattern indicates growing hostility toward media actors, likely reflecting the perception of their role in communicating and legitimizing government vaccination policies.

Finally, the focus of political opposition shifted. While before July 12 users primarily targeted the government at large ("gouvernement"/government, "gouvernants"/rulers, "politiciens"/politicians), after the speech the discourse centered more directly on President Macron as an individual, with terms such as "macronistes" (Macron supporters) and "président" (president) showing significant increases. This highlights how the speech personalized the conflict, transforming generalized anti-government sentiment into a direct confrontation with Macron himself.

In line with prior work on collective action and outgroup salience (Lüders et al. 2021), vertical references to authority figures such as the government and the president were key outgroups. However, our results also highlight a substantial shift in focus: vaccinated individuals (i.e., horizontal comparison) emerged as a highly salient outgroup. The centrality of the authority and vaccinated outgroups can be interpreted through the lens of relative deprivation theory (Crosby 1976). For individuals critical of COVID-19 vaccines and related policies, Macron's speech likely crystallized their growing sense of unfairness about their ingroup's treatment by the authority, relative to the vaccinated outgroup, seen as the privileged group.

Incoming Users and Conformism

We focus on the week of Macron's address on mandatory vaccination and the health pass, which triggered an influx of new users (see Figure 1) to examine how users initiate interactions and adapt to community norms.

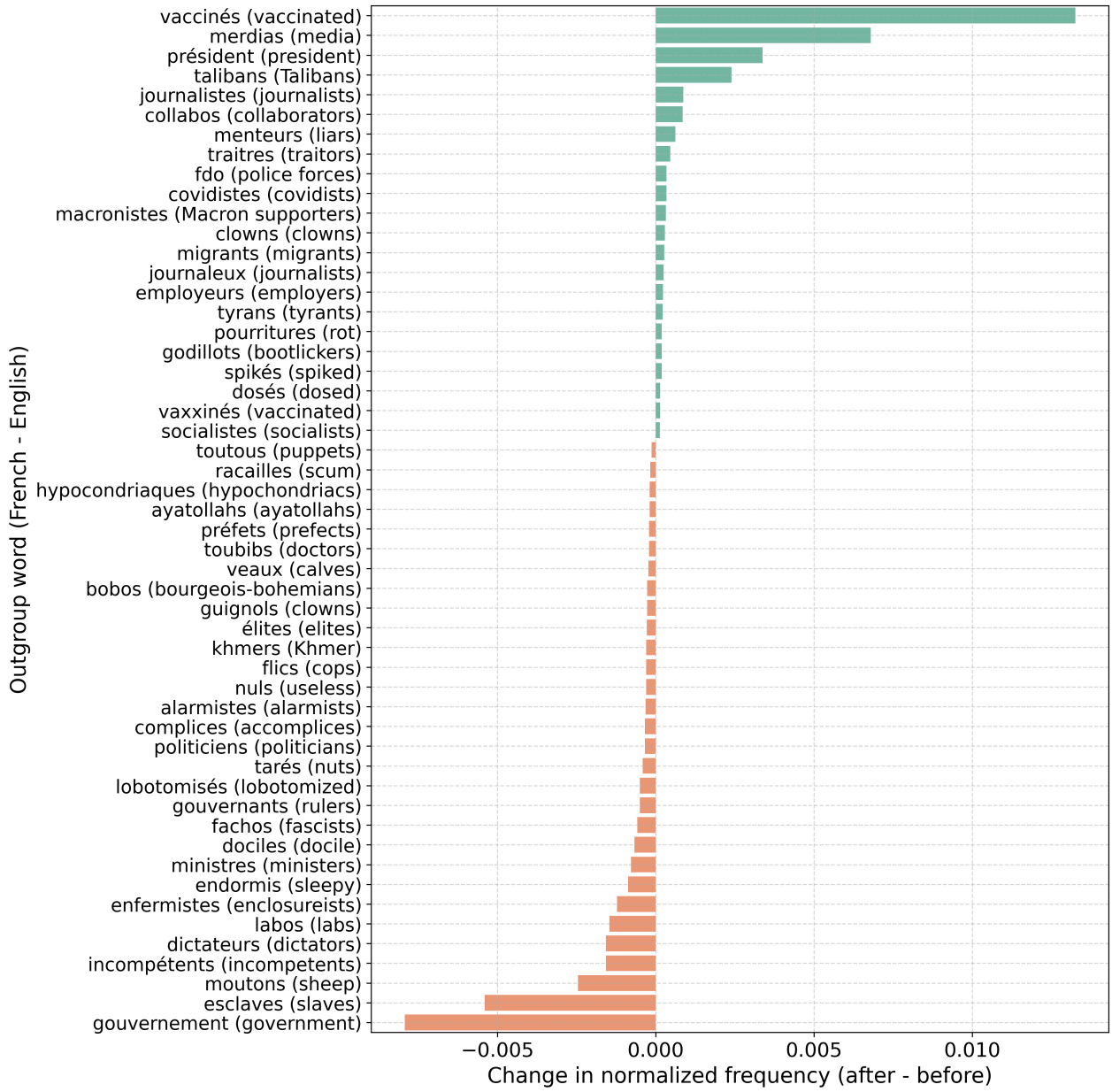


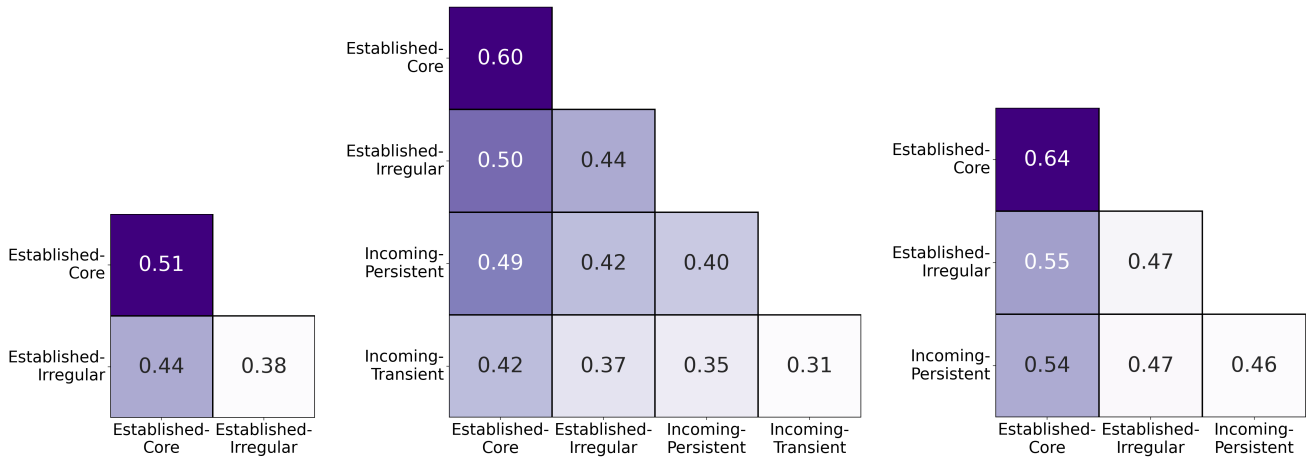
Figure 3: Change in normalized frequency of outgroup labels with statistically significant differences after July 12, 2021, highlighting increased focus on vaccinated individuals, media actors, and President Macron.

Methods

To analyze the linguistic behavior of incoming users and compare it to established users, we group users based on when they began posting and their posting frequency. From the week before Macron's speech to the week after, we categorize users into four distinct groups:

- Established-Core (n=427): The top 10% of users who posted the most (out of all users in the dataset) and were active for at least ten different months, thus having posted before Macron's speech;
- Established-Irregular (n=8931): Users who started posting before Macron's speech but not as often as the Established-Core;
- Incoming-Persistent (n=2068): Users who started posting on the week of Macron's speech and continued after;
- Incoming-Transient (n=3321): Users who started posting on the week of Macron's speech but did not post after.

We use a sentence embedding model pre-trained on French data (Sentence-CamemBERT-base, Tuan 2023) and based on CamemBERT (Martin et al. 2020) and Sentence-BERT (Reimers and Gurevych 2019) to create sentence-embeddings for each tweet. Prior work has also shown



A - Week 2021-27 **B - Week 2021-28** **C - Week 2021-29**

Figure 4: Cosine similarity between the Established-Core, Established-Irregular, Incoming-Persistent, and Incoming-Transient groups. We focused on the week before Macron's speech (A - 2021-27), the week of his speech (B - 2021-28), and the week after (C - 2021-29). The Established-Core consistently show the highest within-group similarity, which increases over time. The Incoming-Transient display the lowest within-group similarity and are the most distinct from the Established-Core. In contrast, the Incoming-Persistent are more similar to the Established-Core.

CamemBERT's effectiveness for French Twitter debates on COVID-19 vaccination, where it reached over 70% accuracy in stance classification and 90% in topic classification (Sauvayre 2023). The sentence embeddings are used to calculate the cosine similarity between tweets within and across user groups. This approach enables us to analyze how closely aligned the language of incoming users is to the one of established ones.

To better understand the distinctiveness of each group, we complement our analysis with Fightin' Words (Monroe, Colaresi, and Quinn 2008), implemented through the Convokit package (Chang et al. 2020), which identifies words or n-grams disproportionately used by one group compared to others.

Results

To better understand the dynamics within and between different groups of users, we measured, for weeks 2021-27 (i.e., the week before Macron's speech), 2021-28 (i.e., the week of his speech), and 2021-29 (i.e., the week after his speech), the cosine similarity between tweets of different user groups (i.e., Established-Core, Established-Irregular, Incoming-Persistent, and Incoming-Transient). Figure 4 shows our results for the different weeks; additional information, including standard deviations and bootstrapped confidence intervals, is provided in the Appendix.

The Established-Core consistently exhibited the highest within-group similarity, which increased steadily over the weeks (.51 in week 2021-27, .60 in week 2021-28, and .64 in week 2021-29). In week 2021-27, the Established-Irregular displayed the lowest within-group similarity (.38) and were more similar to the Established-Core than to themselves (.44). During the week of Macron's speech (2021-28), incoming users, particularly the Incoming-Transient,

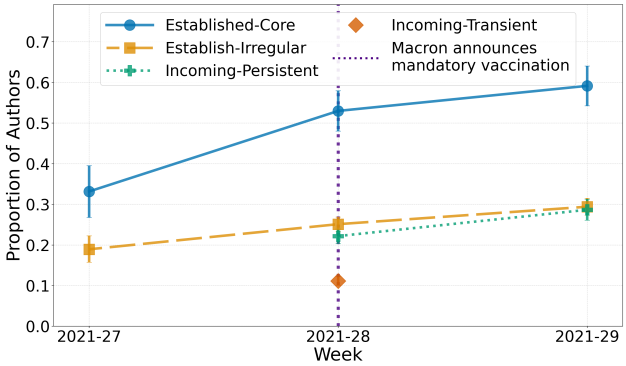


Figure 5: Proportion of outgroup label use by authors in different groups, with 95% confidence interval error bars. The Established-Core use outgroup labels most, the Incoming-Transient least, while the Established-Irregular and Incoming-Persistent show similar patterns.

exhibited the lowest similarity to the Established-Core (.42) and the lowest within-group similarity (.31). The Incoming-Persistent behaved more similarly to the Established-Irregular, and we observed an increase in similarity both within groups (from .44 to .47 for the Established-Irregular and from .40 to .46 for the Incoming-Persistent) and with the Established-Core (from .50 to .55 for the Established-Irregular and from .49 to .54 for the Incoming-Persistent) between weeks 2021-28 and 2021-29.

Additionally, we investigated the different groups' use of pronouns and outgroup labels. We found no substantial differences in pronoun usage; established users shifted from 1SG to 1PL pronouns after Macron's speech, while incoming users entered the discussion with a higher prevalence

of 1PL pronouns (see Appendix for additional detail). Regarding outgroup labels (Figure 5), the Established-Core used them more than other groups, while the Incoming-Persistent and Established-Irregular exhibited similar trends. The Incoming-Transient used less outgroup labels.

Finally, to better understand the speech patterns that distinguish each group, we used the Fightin' Words method (Monroe, Colaresi, and Quinn 2008) to identify the most distinctive ngrams (i.e., continuous sequences of one to three words in our case) used during the week of Macron's speech (see Table 4, Appendix). The primary distinction lied in the focus on political, policy-related, or vaccination-related topics. The Established-Core and Incoming-Persistent were characterized by their criticism of the government and COVID-19 policies. The Established-Core emphasized opposition to the government (e.g., #StopDictatureSanitaire *[stop health dictatorship]*) while the Incoming-Persistent users focused on specific COVID-19 policies (e.g., #PassSanitaireDeLaHonte *[shame pass]*) and broader values (#LiberteEgaliteFraternite *[liberty, equality, fraternity]*). In contrast, the Incoming-Transient focused primarily on vaccination itself (e.g., "vaccin" *[vaccine]*).

Vaccination remained a central issue across all groups. The hashtag #antivax was the most distinctive term for both incoming groups, while #JeNeMeVaccineraiPas *[I won't get vaccinated]* was prominent among Established-Core users. This suggests that while all groups expressed concerns about vaccination, the Established-Core focused more on opposing the government, the Incoming-Persistent emphasized policy critiques, and the Incoming-Transient users directly expressed their concern for vaccination.

Our findings suggest that incoming users who continued posting—likely high identifiers—adhered closely to group norms, as evidenced by their linguistic similarity to Established-Core users, use of 1PL pronouns, and outgroup labels. This alignment may reflect efforts to demonstrate commitment during an uncertain membership phase, consistent with theories of social identification (Klein, Spears, and Reicher 2007; Branscombe et al. 1999). This contrasts with those who posted only briefly—likely low identifiers—, who showed minimal linguistic adaptation.

Discussion and Conclusion

Our findings contribute to research on collective identity by demonstrating how digital traces can reveal processes of identity emergence.

First, we build on prior computational analyses of French vaccine skepticism. Faccin et al. (2022) showed that vaccine-critical Twitter communities formed echo chambers, while Sauvayre (2023) demonstrated that CamemBERT could classify vaccine-related tweets by stance and topic. Extending this work, we examine the linguistic mechanisms—pronoun use, outgroup labeling, and convergence between user groups—through which oppositional identities are discursively constructed and reinforced.

Second, our results highlight the dual effects of health interventions like the health pass or mandatory vaccination. Consistent with previous research showing increased vaccination alongside heightened skepticism (Ward et al. 2022) and perceived rejection catalyzing ingroup cohesion (Branscombe et al. 1999), we find that mandatory measures crystallized resistance, solidifying a previously indistinct vaccine-critical identity with strengthened ingroup cohesion and sharpened outgroup boundaries.

Finally, our study has implications for policymakers and platform regulators. Linguistic markers such as shifts in pronoun usage and the emergence of new outgroup labels can serve as early indicators of polarization. Integrating such signals into monitoring systems may help anticipate the unintended social consequences of public health measures, allowing for more responsive and less polarizing communication strategies.

Limitations and Future Work

Our findings are correlational; while Macron's speech coincided with collective identity crystallization, causality remains unestablished. Future work could use surveys to directly assess user motivations and reactions. Additionally, the use of cosine similarity as a measure of linguistic similarity comes with limitations, as the sentence embeddings are not easily interpretable. We supplemented it with pronoun, outgroup reference, and distinctive words analyses. Future studies could incorporate additional indicators like social network analysis to examine ingroup cohesion over time and sentiment analysis to explore emotional synchronization or well-being effects linked to collective identity (Branscombe, Schmitt, and Harvey 1999). Finally, future research should explore this phenomenon in different contexts to enhance generalizability and deepen theoretical understanding of identity emergence.

Conclusion

Through linguistic analysis of a year of social media discourse criticizing COVID-19 vaccines and related policies in France (e.g., mandatory vaccination, health pass), we gained insights into collective identity formation. We analyzed linguistic patterns—such as pronoun usage, outgroup labeling, and cosine similarity—to trace the evolution of this identity. We found President Macron's speech on mandatory vaccination and the health pass proving to be a pivotal event. This speech marked a shift from first-person singular to first-person plural pronouns and increased focus on vaccinated individuals, the media, and President Macron as outgroups. Additionally, many users joined the conversation during this period, and those who continued posting showed increasing linguistic similarity to established members, integrating further into the community.

Previous research showed that institutionalized rejection based on identity can fuel major social movements (Marx 1998). In France, critics of COVID-19 vaccines may have perceived Macron's announcement as a form of institutionalized rejection, threatening core values like freedom and bodily autonomy. This perceived rejection likely strengthened collective identity among vaccine-hesitant groups. Our findings suggest that mandating vaccination may have had a counterproductive effect, reinforcing the resolve of these groups, making them more vocal and resistant, and potentially deepening societal polarization.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (Grant 18/CRT/6049) and the European Research Council under the EU Horizon 2020 programme (Grant 802421). Dong Nguyen was supported by the "Digital Society – The Informed Citizen" programme partly funded by the Dutch Research Council (Project 410.19.007). We thank Ariana Sepahpour-Fard for annotation validation.

References

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 258–265. IEEE.

Bardosh, K.; De Figueiredo, A.; Gur-Arie, R.; Jamrozik, E.; Doidge, J.; Lemmens, T.; Keshavjee, S.; Graham, J. E.; and Baral, S. 2022. The unintended consequences of COVID-19 vaccine policy: why mandates, passports and restrictions may cause more harm than good. *BMJ Global Health*, 7(5): e008684.

Baumeister, R. F.; and Leary, M. R. 1995. The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, 117(3): 497–529.

Branscombe, N. R.; Ellemers, N.; Spears, R.; and Doosje, B. 1999. The context and content of social identity threat. In Ellemers, N.; Spears, R.; and Doosje, B., eds., *Social identity: Context, commitment, content*, 35–58. New Jersey, NJ: Blackwell Science.

Branscombe, N. R.; Schmitt, M. T.; and Harvey, R. D. 1999. Perceiving pervasive discrimination among African Americans: Implications for group identification and well-being. *Journal of personality and social psychology*, 77(1): 135.

Brewer, M. B. 1991. The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5): 475–482.

Brewer, M. B.; and Gardner, W. 1996. Who is this" We"? Levels of collective identity and self representations. *Journal of personality and social psychology*, 71(1): 83.

Bronner, L.; and Mandard, S. 2021. À Paris, le noyau dur et hétérogène de la contestation contre le passe sanitaire. *Le Monde*.

Chang, J. P.; Chiam, C.; Fu, L.; Wang, A.; Zhang, J.; and Danescu-Niculescu-Mizil, C. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 57–60. 1st virtual meeting: Association for Computational Linguistics.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini, A. 2021. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 89–96.

Crosby, F. 1976. A model of egoistical relative deprivation. *Psychological review*, 83(2): 85.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 307–318. New York, NY, USA: Association for Computing Machinery.

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, 273–274.

Drury, J.; and Reicher, S. 2000. Collective action and psychological change: The emergence of new social identities. *British journal of social psychology*, 39(4): 579–604.

Faccin, M.; Gargiulo, F.; Atlani-Duault, L.; and Ward, J. K. 2022. Assessing the influence of French vaccine critics during the two first years of the COVID-19 pandemic. *PLoS One*, 17(8): e0271157.

Flesher Fominaya, C. 2010. Collective identity in social movements: Central concepts and debates. *Sociology compass*, 4(6): 393–404.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gable, J. S.; Sauvayre, R.; and Chauvière, C. 2023. Fight against the mandatory COVID-19 immunity passport on twitter: natural language processing study. *Journal of Medical Internet Research*, 25: e49435.

Gallup. 2019. Wellcome Global Monitor 2018: First Wave Findings. Technical report, Wellcome Trust.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

González-Padilla, D. A.; and Tortolero-Blanco, L. 2020. Social media influence in the COVID-19 Pandemic. *International braz j urol*, 46(Suppl 1): 120–124.

Guimond, S.; and Dambrun, M. 2002. When prosperity breeds intergroup hostility: The effects of relative deprivation and relative gratification on prejudice. *Personality and social psychology bulletin*, 28(7): 900–912.

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–25.

Hossain, T.; Logan IV, R. L.; Ugarte, A.; Matsubara, Y.; Young, S.; and Singh, S. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics.

Kadić-Maglajlić, S.; Lages, C. R.; and Pantano, E. 2024. No time to lie: Examining the identity of pro-vaccination and anti-vaccination supporters through user-generated content. *Social Science & Medicine*, 347: 116721.

Khazraee, E.; and Novak, A. N. 2018. Digitally mediated protest: Social media affordances for collective identity construction. *Social Media+ Society*, 4(1): 2056305118765740.

Klein, O.; Spears, R.; and Reicher, S. 2007. Social identity performance: Extending the strategic side of SIDE. *Personality and social psychology review*, 11(1): 28–45.

Labov, W. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. New Jersey, NJ: John Wiley & Sons.

Landis, J. R.; and Koch, G. G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Lee, S. H.; Tak, J.-Y.; Kwak, E.-J.; and Lim, T. Y. 2020. Fandom, social media, and identity work: The emergence of virtual community through the pronoun "we". *Psychology of Popular Media*, 9(4): 436.

Linabary, J. R.; Corple, D. J.; and Cooky, C. 2020. Feminist activism in digital space: Postfeminist contradictions in# WhyIStayed. *New Media & Society*, 22(10): 1827–1848.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Livingstone, A. G.; Haslam, S. A.; Postmes, T.; and Jetten, J. 2011. "We Are, Therefore We Should": Evidence That In-Group Identification Mediates the Acquisition of In-Group Norms. *Journal of Applied Social Psychology*, 41(8): 1857–1876.

Lüders, A.; Dinkelberg, A.; and Quayle, M. 2022. Becoming "us" in digital spaces: How online users creatively and strategically exploit social media affordances to build up social identity. *Acta Psychologica*, 228: 103643.

Lüders, A.; Urbanska, K.; Wollast, R.; Nugier, A.; and Guimond, S. 2021. Bottom-up populism: How relative deprivation and populist attitudes mobilize leaderless anti-government protest. *Journal of Social and Political Psychology*, 9(2): 506–519.

Maitland, K.; and Wilson, J. 1987. Pronominal selection and ideological conflict. *Journal of pragmatics*, 11(4): 495–512.

Martin, L.; Muller, B.; Ortiz Suárez, P. J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. Online: Association for Computational Linguistics.

Marx, A. W. 1998. *Making race and nation: A comparison of South Africa, the United States, and Brazil*. Cambridge, UK: Cambridge University Press.

Melucci, A. 1980. The new social movements: A theoretical approach. *Social science information*, 19(2): 199–226.

Melucci, A. 1985. The Symbolic Challenge of Contemporary Movements. *Social Research*, 52(4): 789–816.

Melucci, A. 1988. Getting involved: identity and mobilization in social movements. *International social movement research*, 1(26): 329–48.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.

Moreland, R. L. 1985. Social categorization and the assimilation of" new" group members. *Journal of Personality and Social Psychology*, 48(5): 1173.

Morselli, D.; Beramendi, M.; Bendali, Z.; and Fillieule, O. 2023. A Longitudinal Approach to Online "Collective Identity Work": The Case of the Gilets Jaunes in the Var Department. *Mobilization: An International Quarterly*, 28(3): 301–320.

Motta, M.; Callaghan, T.; Sylvester, S.; and Lunz-Trujillo, K. 2023. Identifying the prevalence, correlates, and policy consequences of anti-vaccine social identity. *Politics, Groups, and Identities*, 11(1): 108–122.

Nguyen, D.; and P. Rosé, C. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 76–85. Portland, Oregon: Association for Computational Linguistics.

Noel, J. G.; Wann, D. L.; and Branscombe, N. R. 1995. Peripheral ingroup membership status and public negativity toward outgroups. *Journal of personality and social psychology*, 68(1): 127.

O'Reilly, C.; Mannion, S.; Maher, P. J.; Smith, E. M.; MacCarron, P.; and Quayle, M. 2024. Strategic attitude expressions as identity performance and identity creation in interaction. *Communications Psychology*, 2(1): 27.

Papapavlou, A.; and Sophocleous, A. 2009. Relational social deixis and the linguistic construction of identity. *International Journal of Multilingualism*, 6(1): 1–16.

Pennebaker, J. W. 2011. The secret life of pronouns. *New Scientist*, 211(2828): 42–45.

Pennebaker, J. W.; and Chung, C. K. 2014. Counting little words in big data: The psychology of individuals, communities, culture, and history. In Forgas, J. P.; Vincze, O.; and László, J., eds., *Social cognition and communication*, 25–42. New York, NY: Psychology Press.

Pera, A.; and Aiello, L. M. 2024. Narratives of collective action in youtube's discourse on veganism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1220–1236.

Peretti-Watel, P.; Seror, V.; Cortaredona, S.; Launay, O.; Raude, J.; Verger, P.; Fressard, L.; Beck, F.; Legleye, S.; L'Haridon, O.; et al. 2020. A future vaccination campaign against COVID-19 at risk of vaccine hesitancy and politicisation. *The Lancet infectious diseases*, 20(7): 769–770.

Pickett, C. L.; and Leonardelli, G. J. 2006. Using collective identities for assimilation and differentiation. In Postmes, T.; and Jetten, J., eds., *Individuality and the group: Advances in social identity*, 56–73. California, CA: Sage Publications.

Polletta, F.; and Jasper, J. M. 2001. Collective identity and social movements. *Annual review of Sociology*, 27(1): 283–305.

Reavley, N. J.; and Pilkington, P. D. 2014. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, 2: e647.

Reicher, S. D. 1996. 'The Battle of Westminster': Developing the social identity model of crowd behaviour in order to explain the initiation and development of collective conflict. *European journal of social psychology*, 26(1): 115–134.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Rousseau, D.; and Van Der Veen, A. M. 2005. The emergence of a shared identity: An agent-based computer simulation of idea diffusion. *Journal of Conflict Resolution*, 49(5): 686–712.

Sauvayre, R. 2023. Obligation vaccinale : quand l'histoire nous invite à réfléchir sur le présent. In Israel-Jost, V.; and Weil-Dubuc, P.-L., eds., *Éthique vaccinale : Ce que nous a appris la crise sanitaire*, 145–156. Toulouse: Érès.

Sauvayre, R.; Vernier, J.; and Chauvière, C. 2022. An Analysis of French-Language Tweets About COVID-19 Vaccines: Supervised Learning Approach. *JMIR Medical Informatics*, 10(5): e37831.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2725–2732.

Sepahpour-Fard, M.; and Quayle, M. 2022. How Do Mothers and Fathers Talk About Parenting to Different Audiences? Stereotypes and Audience Effects: An Analysis of r/Daddit, r/Mommit, and r/Parenting Using Topic Modelling. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2696–2706. New York, NY, USA: Association for Computing Machinery.

Sepahpour-Fard, M.; Quayle, M.; Schuld, M.; and Yasseri, T. 2023. Using word embeddings to analyse audience effects and individual differences in parenting Subreddits. *EPJ Data Science*, 12(1): 38.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1): 4787.

Simon, B.; and Hamilton, D. L. 1994. Self-stereotyping and social context: the effects of relative in-group size and in-group status. *Journal of personality and social psychology*, 66(4): 699.

Simon, B.; Loewy, M.; Stürmer, S.; Weber, U.; Freytag, P.; Habig, C.; Kampmeier, C.; and Spahlinger, P. 1998. Collective identification and social movement participation. *Journal of personality and social psychology*, 74(3): 646.

Smith, L. G.; Gavin, J.; and Sharp, E. 2015. Social identity formation during the emergence of the occupy movement. *European Journal of Social Psychology*, 45(7): 818–832.

Smithey, L. A. 2009. Social movement strategy, tactics, and collective identity 1. *Sociology Compass*, 3(4): 658–671.

Snow, D. 2001. *Collective Identity and Expressive Forms*, 2212–2219. Amsterdam, NL: Elsevier.

Spann, B.; Agarwal, N.; Stafford, D.; and Okeke, O. 2023. Evaluating the Emergence of Collective Identity using Socio-Computational Techniques. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, 1389–1394. Association for Computing Machinery.

Tajfel, H.; and Turner, J. C. 1979. An Integrative Theory of Intergroup Conflict. *The Social Psychology of Intergroup Relations*, 33(47): 74.

Taylor, D. M.; and Dube, L. 1986. Two faces of identity: The "I" and the "We". *Journal of Social Issues*, 42(2): 81–98.

Taylor, V.; and Whittier, N. E. 1992. Collective identity in social movement communities: Lesbian feminist mobilization. *Frontiers in social movement theory*, 104: 109–21.

Tuan, D. V. 2023. sentence-camembert-base. Hugging Face model repository, accessed 2024-07-19.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Technical report, SSRN.

Turner, J. C.; Oakes, P. J.; Haslam, S. A.; and McGarty, C. 1994. Self and collective: Cognition and social context. *Personality and social psychology bulletin*, 20(5): 454–463.

Van Stekelenburg, J. 2013. Collective identity. In *The Wiley-Blackwell encyclopedia of social and political movements*, 219–225. New Jersey, NJ: Wiley-Blackwell.

van Zomeren, M.; Kutlaca, M.; and Turner-Zwinkels, F. 2018. Integrating who "we" are with what "we"(will not) stand for: A further extension of the Social Identity Model of Collective Action. *European Review of Social Psychology*, 29(1): 122–160.

Van Zomeren, M.; Postmes, T.; and Spears, R. 2012. On conviction's collective consequences: Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology*, 51(1): 52–71.

Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 280–289.

Ward, J. K.; Gauna, F.; Gagneux-Brunon, A.; Botelho-Nevers, E.; Cracowski, J.-L.; Khouri, C.; Launay, O.; Verger, P.; and Peretti-Watel, P. 2022. The French health pass holds lessons for mandatory COVID-19 vaccination. *Nature Medicine*, 28(2): 232–235.

Yardi, S.; and Boyd, D. 2010. Tweeting from the Town Square: Measuring Geographic Local Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1): 194–201.

Yoder, M.; Perry, C.; Brown, D.; Carley, K.; and Pruden, M. 2023. Identity Construction in a Misogynist Incels Forum. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, 1–13. Association for Computational Linguistics.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. The study analyzes public tweets and investigates group-level linguistic patterns without profiling individuals or revealing personal data.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. The abstract and introduction reflect the study's focus on the emergence of collective identity among vaccine-critical Twitter users, highlighting linguistic shifts and group dynamics.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we demonstrate the appropriateness of our methods by showing their validation in prior published work, and by using multiple complementary approaches to strengthen our claims.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. First, we filtered out tweets that were not aligned with vaccine criticism by training a classifier on manually labeled data. This helped reduce noise from pro-vaccine users and unrelated content. Second, we examined the role of hashtags in shaping linguistic trends, particularly the shift from individual (1SG) to collective (1PL) pronouns. To test for artifacts, we re-ran our analyses excluding hashtags used for data collection and found consistent patterns, suggesting that our results were not merely artifacts of the data collection method or specific hashtags.
   (e) Did you describe the limitations of your work? Yes. We note that the findings are correlational and acknowledge the limitations of using cosine similarity on sentence embeddings. We suggest survey methods and sentiment or network analysis as future directions.
   (f) Did you discuss any potential negative societal impacts of your work? No. While we do not anticipate any direct negative societal impacts from our work, we recognize the importance of considering such effects. Our analysis is focused on understanding the emergence of collective identity in response to vaccination policies, with the aim of helping policymakers better understand public reactions to health mandates. By shedding light on how policy announcements may unintentionally reinforce resistance, our goal is to inform more effective, inclusive, and context-sensitive public health strategies in future crises such as COVID-19.
   (g) Did you discuss any potential misuse of your work? No. We do not foresee any significant potential for misuse of our work. The study focuses on aggregated linguistic patterns and group-level behaviors without identifying individuals. All data analyzed consists of public tweets, and no personally identifiable information is disclosed. Since the goal is to understand collective identity formation in response to public policy, we believe the work is unlikely to be misused for harmful purposes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No. We plan to make anonymized data available upon request to ensure responsible data sharing. While the dataset consists of public tweets, we take precautions to protect user privacy by excluding any personally identifiable information. Additionally, we will release the code and documentation necessary to reproduce our findings.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Not applicable.
   (b) Have you provided justifications for all theoretical results? Not applicable.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not applicable.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not applicable.
   (e) Did you address potential biases or limitations in your theoretical framework? Not applicable.
   (f) Have you related your theoretical results to the existing literature in social science? Not applicable.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Not applicable.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Not applicable.
   (b) Did you include complete proofs of all theoretical results? Not applicable.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, but as stated in the paper, we will make the code used available upon publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. We included the key training details such as data splits, model architecture, and evaluation metrics within the constraints of the paper's length. Additional details, including full hyperparameter settings and training procedures, will be made available alongside the released code to ensure full reproducibility.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. We employed statistical significance testing throughout the study and reported p-values and 95% confidence intervals where appropriate, particularly in our analyses of pronoun usage and outgroup labeling. These measures help assess the robustness and reliability of our findings.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No. We used a combination of Google Colab and a local workstation equipped with a GPU to run our experiments, including training the classifier and computing embeddings. While we did not include detailed compute specifications in the paper due to space constraints, this information will be provided alongside the released code and documentation.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. We justify the appropriateness of our evaluation approach across all methods used. For the classifier, we report standard performance metrics (precision, recall, F1), demonstrating its suitability for filtering vaccine-critical tweets. For linguistic analyses, we use proportion z-tests and report significance values and confidence intervals to support our claims. We also measure cosine similarity using sentence embeddings and apply the Fightin' Words method to identify distinctive language features. All these approaches are grounded in prior work, and we draw on established theories and findings to guide interpretation of the results.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No. We did not explicitly discuss the cost of misclassification because the classifier was used primarily as a filtering tool to exclude irrelevant tweets (e.g., pro-vaccine or unrelated content), rather than for downstream tasks requiring high-stakes decision-making. Given the high precision and recall of the classifier, and the aggregate nature of our linguistic analyses, we believe minor misclassifications are unlikely to significantly affect our findings or conclusions.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. CamemBERT, Convokit, and other models and tools are cited.

(b) Did you mention the license of the assets? We did not explicitly mention license information in the paper due to space constraints, but we only used publicly available tools and models (e.g., CamemBERT, ConvoKit, Sentence-Transformers) that are released under permissive open-source licenses. License information can be provided alongside the released code and documentation.

(c) Did you include any new assets in the supplemental material or as a URL? No. The code will be released upon publication.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, we did not discuss consent as the study uses publicly available Twitter data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No. The study uses publicly available Twitter data.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? We did not explicitly discuss the FAIR principles in the paper due to space constraints. However, we are committed to ensuring our curated dataset adheres to these principles upon release. Specifically, we plan to provide clear documentation, anonymized data, and standardized formats. Access will be granted upon request to manage ethical considerations.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? We did not include a datasheet for the dataset in the current submission due to space constraints and because the dataset is not publicly released. However, we recognize the importance of datasheets for transparency and responsible data sharing. We plan to provide a complete datasheet upon request.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Not applicable.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable.

(d) Did you discuss how data is stored, shared, and deidentified? Not applicable.

Classifier

To prepare the data for our analyses, we built a classifier to identify tweets critical of COVID-19 vaccines and related policies, such as the health pass. The codebook underlying this classifier was developed through a pilot annotation process in which we iteratively refined categories until they reliably captured the relevant discourse. This process resulted in four categories of included tweets and four categories of excluded tweets, which were then used to annotate the final sample for training the classifier.

Categories of included tweets:

- Anti-vaccination: Tweets clearly expressing opposition to the COVID-19 vaccine;

- Anti-policy: Tweets criticizing government policies related to vaccination without expressing a direct position on vaccination itself;
- Vaccination status with opposition: Tweets stating whether the user is vaccinated or not, while expressing opposition to vaccine-related or government policies;
- Antivax hashtags: Tweets containing hashtags, the majority of which are linked to antivax movements.

Examples of included tweets are provided in Table 2.
Categories of excluded tweets:

- Pro-vaccination: Tweets explicitly in favor of vaccination or criticizing vaccine hesitancy;
- Not individual users: Tweets that do not appear to represent individual users;
- Criticizing both sides: Users criticizing both pro- and anti-vaccine positions;
- Unrelated tweets: Tweets that use at least one of the collection hashtags but whose content is unrelated to COVID-19 or vaccination.

Examples of excluded tweets are provided in Table 1.

Then, we sampled 1% of the dataset (4019 tweets) for manual labeling. As participation fluctuates substantially depending on the period and external circumstances, we sampled tweets proportionally to the number of tweets published each week. In the sample used for manual classification, we identified 3,417 tweets relevant for our analysis. An external collaborator labeled 100 randomly selected tweets, and the inter-annotator agreement with our labels resulted in a Cohen's Kappa score of 0.69, indicating substantial agreement (Landis and Koch 1977).

Given that our data is in French, we fine-tuned CamemBERT for sequence classification (Martin et al. 2020), which is based on the RoBERTa architecture (Liu et al. 2019) and trained on French data, to automatically classify the entire dataset. The labeled data was partitioned into two subsets: 3,500 tweets for training and validation (80% of the total dataset, stratified proportionally to class observations, with 20% used for validation) and 519 tweets for testing. The model was fine-tuned for ten epochs with a batch size of 16. During fine-tuning, we optimized hyperparameters using the validation set. The final model performance is presented in Table 3, showing detailed metrics. Notably, class 1 (i.e., tweets related to the criticism of COVID-19 vaccines and related policies) was well classified, achieving a precision, recall, and F1 score of .97, .96, and .96, respectively.

Bot Detection

In order to asses the potential impact bots have on our analyses and results, we used the Botometer (Sayyadiharikandeh et al. 2020; Varol et al. 2017) to estimate the proportion of bots in our results. We found 6.07% of accounts with a bot score > 0.5, which contributed to 2.35% of tweets in our data. Figure 6 shows the distribution of bot scores. To further investigate their influence, we ran again our main analyses for pronoun use, outgroup labels, and the different categories of users after removing potential bots (i.e., bot score > 0.5).

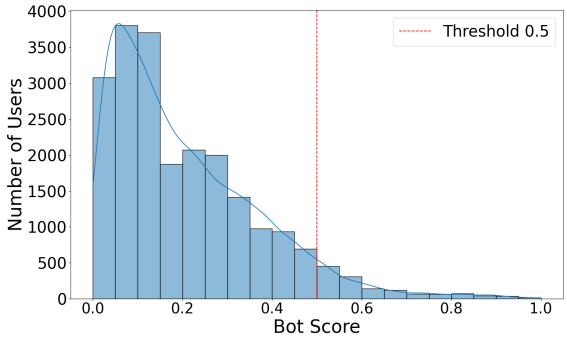


Figure 6: Distribution of bot scores in the dataset. Only a small proportion (6.07%) was found to have a score > 0.5

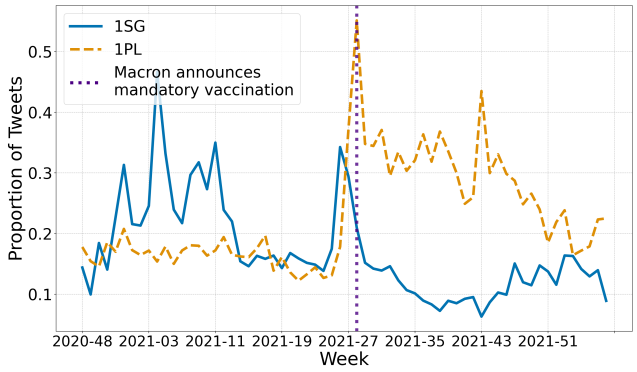


Figure 7: Prevalence of 1SG and 1PL pronouns across weeks, after removing potential bot accounts.

Pronoun Use

Figure 7 shows the prevalence of 1SG and 1PL pronouns over time after removing potential bots. We found that the shift from 1SG to 1PL pronouns persisted, reflecting the results shown in the main paper.

Outgroup Labels

We ran again our analysis on the change of outgroup labels before and after President Macron's speech. Figure 8 shows the change in frequency of outgroup labels when removing potential bots. We found 50 words to significantly change frequency (52 in the analyses including potential bots). However, we observe very similar patterns with tweets showing an increased focus on outgroup labels related to vaccinated people ("vaccinés"/vaccinated, "spikés"/spiked, "dosés"/dosed, and "vaxxinés"/vaccinated), media ("merdias"/a derogatory portmanteau meaning "shitty media", "journalistes"/journalists, and "journaleux"/pejorative term for journalists), and President Macron ("macronistes"/Macron supporters and "président"/president).

Incoming Members

Finally, we ran again the analyses on the four categories of accounts around Macron's speech: Established-Core,

Table 1: Examples of excluded tweets.

Category	Tweet	Tweet Translation
Pro-vaccination	#DictatureSanitaire Avant d'utiliser les mots, réfléchir permet de ne pas se tromper. Voilà le résultat d'une vraie dictature. Croyez vous vraiment vivre ça actuellement quitte à minimiser le passé !	#HealthDictatorship Before using words, thinking allows us to avoid mistakes. This is the result of a real dictatorship. Do you really think you are living that right now, at the risk of minimizing the past!
Pro-vaccination	#C'est quoi la prochaine étape de ces déchets ? Tabasser ceux qui se font vacciner ?? Quelle honte ces nazes ! #antivax #GiletsJaunes #VaccinezVous #COVID19 #VariantDelta	#What's the next step for this trash? Beating up those who get vaccinated?? What a shame these idiots! #antivax #YellowVests #GetVaccinated #COVID19 #DeltaVariant
Not individual users	#Passsanitaire : les #manifestations ont débuté dans de nombreuses villes en #France #Covid #antivaccin #obligationvaccinale	#HealthPass: #protests have started in many cities in #France #Covid #antivax #mandatoryvaccination
Not individual users	LES INDISPENSABLES #Antivax – Jusqu'où ira Didier Raoult ? Analyses et décryptages de @fanny_weil dans Les Indispensables #24hPujadas #LCI #La26	THE ESSENTIALS #Antivax – How far will Didier Raoult go? Analysis and decryption by @fanny_weil in Les Indispensables #24hPujadas #LCI #La26
Criticising both sides	C quand la fin de la guerre des #Provax et des #antivax ?	When is the end of the war between #Provax and #antivax?
Criticising both sides	Oups... #Macron @EmmanuelMacron (VS) #antivax Les non-vaccinés...	Oops... #Macron @EmmanuelMacron (VS) #antivax The unvaccinated...
Unrelated tweets	Marine Le Pen veut inscrire la "priorité nationale" dans la Constitution via un référendum et l'appliquer au logement social Via @ActusMondial #Terrorisme #Resistance #AFP @GJaunes #SousMarinGate #NonAuPassDeLaHonte	Marine Le Pen wants to include "national priority" in the Constitution via a referendum and apply it to social housing Via @ActusMondial #Terrorism #Resistance #AFP @YellowVests #SubmarineGate #NoToTheShamePass
Unrelated tweets	Al-Qaïda et l'État islamique : trois différences majeures qui distinguent les deux organisations terroristes Via/ @ActusMondial #AFP #NonAuPassDeLaHonte #Terrorisme	Al-Qaeda and the Islamic State: three major differences that distinguish the two terrorist organizations Via/ @ActusMondial #AFP #NoToTheShamePass #Terrorism

Established-Irregular, Incoming-Persistent, and Incoming-Transient users. Figure 9 shows the mean pairwise cosine similarity between the categories of accounts, after removing bots. We found no substantial change from the dynamics observed in the main analyses. Similarly, regarding the prevalence of outgroup labels across weeks in different categories (Figure 10), our findings confirmed the negligible influence of bots on the main results of the paper.

Additional Analyses for Pronoun Use

This section presents additional analyses related to pronoun use. Figure 12 complements Figure 2, directly comparing the use of 1SG and 1PL pronouns before and after Macron's speech. Furthermore, we examine the influence of hashtags in shaping pronoun usage patterns, and we compare the pronoun use of different user groups.

Hashtags and Pronoun Use

In the main analyses, we counted personal pronouns (i.e., "I" or "we") when they appeared at the onset of a hashtag

Table 2: Examples of included tweets classified into four categories.

Category	Tweet	Tweet Translation
Anti-vaccination	Croyez le bien que j'vais me faire vacciner avec un truc développé en 6 mois qui va nous faire pousser 3 têtes dans 10 ans juste pour aller bouffer dans un restau ou aller au cinéma. On a un sacré problème en France. #DictatureSanitaire #Castex	Believe me, I'm going to get vaccinated with something developed in 6 months that will make us grow 3 heads in 10 years just to eat in a restaurant or go to the movies. We have a serious problem in France. #HealthDictatorship #Castex
Anti-policy	Que ceux qui veulent rester cloitré restent chez eux ! Que les autres vivent libres ! #DictatureSanitaire #confinement	Those who want to stay locked up should stay at home! Let the others live free! #HealthDictatorship #lockdown
Vaccination status with opposition	Chacun est libre de choisir s'il souhaite se faire ou non vacciné! Je ne suis en aucun cas un antivax mais contre le pass sanitaire #NonAuPassDeLaHonte	Everyone is free to choose whether or not to get vaccinated! I am by no means an antivax but I am against the health pass #NoToTheShamePass
Antivax hashtags	#NousSommesDesMillions https://t.co/336lMnxiiH	#WeAreMillions https://t.co/336lMnxiiH
Antivax hashtags	#31juillet2021 #NonAuPassDeLaHonte #NonAuPasseDeLaHonte #TousUnis #NousSavons #NousRésistons #NousSommesDesMillions #BoycottDictatureSanitaire https://t.co/Qwkv0aAHZc	#31July2021 #NoToTheShamePass #NoToTheShamePass #AllUnited #WeKnow #WeResist #WeAreMillions #BoycottHealthDictatorship https://t.co/Qwkv0aAHZc

Table 3: Metrics for each class, weighted average, macro average, and accuracy. Class 1 refers to the prediction of tweets related to the criticism of COVID-19 vaccines and related policies, and Class 0 refers to tweets that do not align with our study focus.

	Class 0	Class 1	Weighted Average	Macro Average
Precision	0.783133	0.970183	0.942072	0.876658
Recall	0.833333	0.959184	0.940270	0.896259
F1 Score	0.807453	0.964652	0.941027	0.886053
Accuracy			0.940270	

(e.g., #NousSommesDesMillions *[We are millions]* or #JeNeMeVaccineraiPas *[I won't get vaccinated]*). We extend this analysis by examining the influence of hashtags on pronoun use. Figure 11 shows pronoun use when (A) excluding hashtags; (B) including only hashtags, i.e., removing the body of tweets except for hashtags; and (C) including only the hashtags that we did *not* use for data collection.

Our results indicated that hashtags played a crucial role in pronoun usage, particularly in the observed shift from first-person singular (1PL) to first-person plural (1PL) pronouns. Without hashtags, there was no discernible shift, as 1PL pronouns consistently outnumbered 1SG pronouns. However, when analyzing only hashtags (B), we saw the shift: 1SG pronouns were more prevalent before Macron's speech, while 1PL pronouns dominated afterward. To ensure that the hashtags used in our data collection were not solely responsible for this shift, we excluded them (C) and found a similar pattern, reinforcing our confidence in the results of the main analyses. In summary, our findings highlighted a shift in pronoun use following Macron's speech and underscored the importance of hashtags in this phenomenon.

Pronoun Use in Established and Incoming Users

To compare different groups of users, in particular established (i.e., users who started posting before Macron's speech) and incoming (i.e., users who started posting on

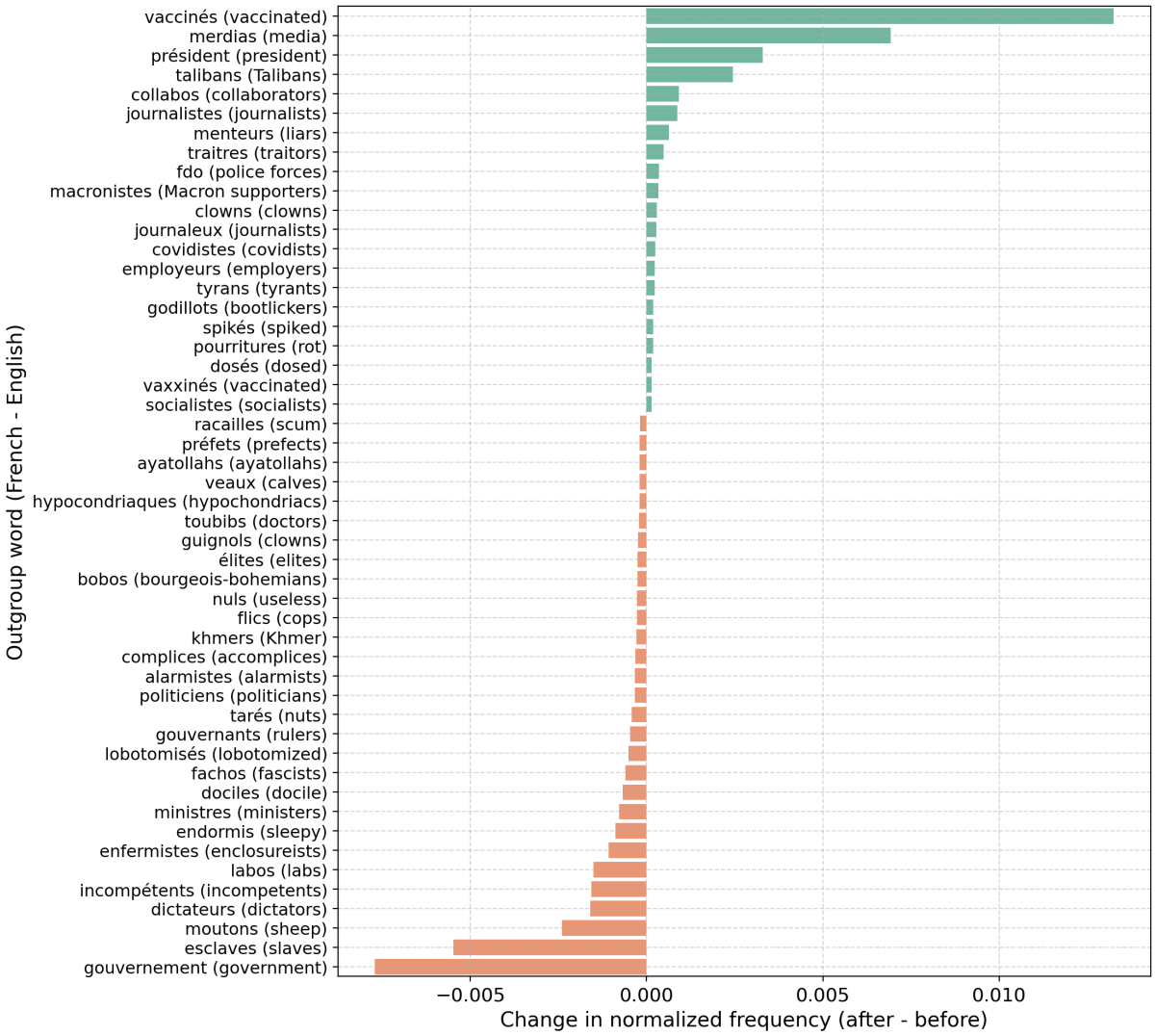


Figure 8: Change in normalized frequency of outgroup labels after July 12, 2021, excluding potential bot accounts.

the week of Macron's speech), we compared their pronoun use. Figure 13 shows the evolution of pronoun use for the four groups of users: (A) Established-Core, (B) Established-Irregular, (C) Incoming-Persistent, and (D) Incoming-Transient. The Established-Core and -Irregular shifted from a first-person singular (1SG) pronoun focus to a first-person plural (1PL) pronoun focus, especially visible after Macron's speech on the week 2021-28. The clearest shift from 1SG to 1PL could be observed in Established-Core users. Similarly, incoming users used more 1PL than 1SG pronouns, as they started posting on the week of Macron's speech. The high use of 1PL pronouns in Incoming-Transient users might be due to the fact that the hashtag #NousSommesDesMillions *[We are millions]* was the second most distinctive word of the group (see Table 4). Additionally, the Incoming-Persistent seemed to maintain a consistently high use of 1PL pronouns compared to established users.

Distinctive Words in Established and Incoming Users

Fighin' Words (Monroe, Colaresi, and Quinn 2008) uses Bayesian techniques to adjust the estimated differences in word usage between groups, ensuring more reliable comparisons by pulling extreme values toward more reasonable estimates (shrinkage) and preventing the model from being influenced by rare or noisy data (regularization).

Table 4 shows the most distinctive words for each group comparison. Z-scores indicate the importance of differences in word usage between groups, with higher absolute scores reflecting greater differences.

Outgroup Labels

To complement the results presented in the main analyses, we present the full list of labels and their translations:

"gouvernement" (government), "moutons" (sheep),

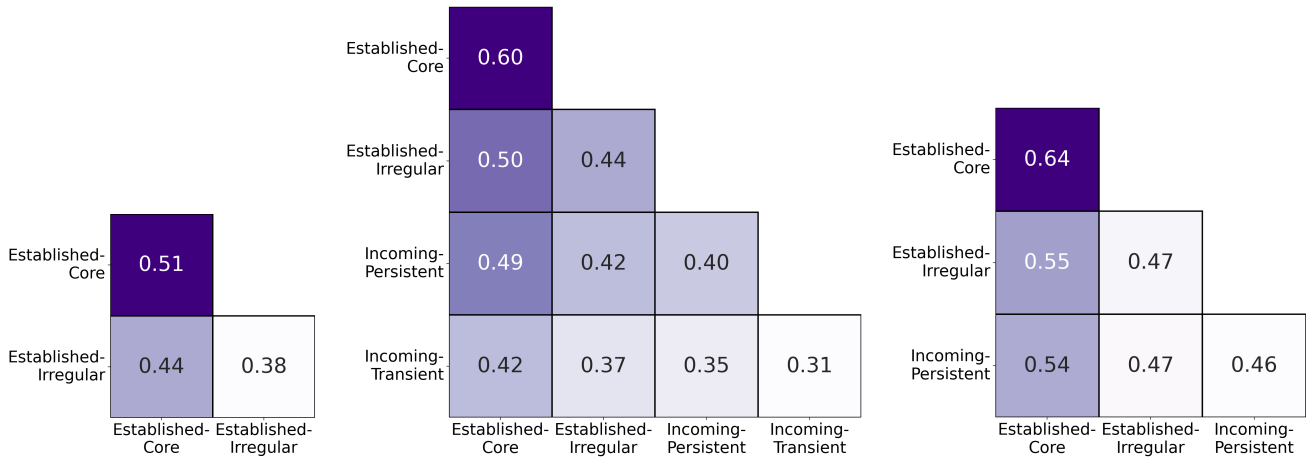


Figure 9: Cosine similarity between the Established-Core, Established-Irregular, Incoming-Persistent, and Incoming-Transient groups, after removing potential bots. We focused on the week before Macron's speech (A - 2021-27), the week of his speech (B - 2021-28), and the week after (C - 2021-29).

Table 4: Most distinctive words and Z-scores comparing different user groups. The words characteristic of group (1) in the comparison are displayed in column (1) while those for group (2) are in column (2). The comparisons suggest that the Established-Core and Incoming-Persistent frame the issue more politically, while the Incoming-Transient focus more on vaccination.

Comparison	**(1)**	**(2)**
Established-Core (1) vs. Incoming-Persistent (2)	#JeNeMeVaccineraiPas (24.39) *[I won't get vaccinated]*, #GiletsJaunes (19.77) *[yellow vests]*, #resistance (19.12), #BoycottPassSanitaire (18.64) *[boycott health pass]*, #SoutienAuxSoignants (16.16) *[support for healthcare workers]*	#antivax (-16.47), #NonAuPassDeLaHonte (-16.10) *[no to the shame pass]*, #LiberteEgaliteFraternite (-11.00) *[liberty equality fraternity]*, medias (-10.69), #PassSanitaireDeLaHonte (-10.49) *[shame health pass]*
Established-Core (1) vs. Incoming-Transient (2)	#JeNeMeVaccineraiPas (16.84), #resistance (16.38), #GiletsJaunes (14.70), #StopDictatureSanitaire (13.56) *[stop health dictatorship]*, #BoycottPassSanitaire (12.65)	#antivax (-21.05), #NousSommesDesMillions (-20.29) *[we are millions]*, vaccin (-15.94) *[vaccine]*, vacciner (-13.81) *[to vaccinate]*, faire (-13.56) *[to do]*

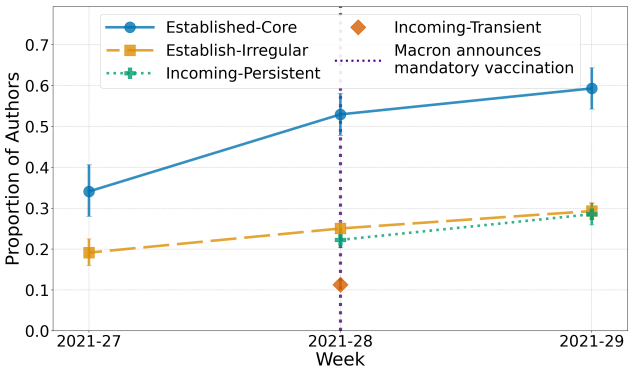


Figure 10: Proportion of outgroup label use by authors in different groups after removing potential bots, with 95% confidence interval error bars.

"flics" (cops), "covidistes" (covidists), "élites" (elites), "vaccinés" (vaccinated), "médias" (media), "fdo" (police forces), "lobotomisés" (lobotomized), "dirigeants" (leaders), "préfets" (prefects), "boomers" (boomers), "macronistes" (Macron supporters), "esclaves" (slaves), "politiciens" (politicians), "gouvernants" (rulers), "tyrans" (tyrants), "dictateurs" (dictators), "terroristes sanitaires" (health terrorists), "politicards" (political hacks), "vendus" (sellouts), "larbins" (minions), "lâches" (cowards), "enfermistes" (enclosureists), "tarés" (nuts), "escrocs" (scammers), "collabos" (collaborators), "charlatans" (charlatans), "crapules" (scoundrels), "mougeons" (maggots), "décérébrés" (brainless), "nantis" (well-off), "fachos" (fascists), "dociles" (docile), "journalistes" (journalists), "naïfs" (naive), "clowns" (clowns), "hypocondriaques" (hypochondriacs), "merdias" (media), "président" (president), "ministres" (ministers), "alarmistes" (alarmists), "bobos" (bourgeois-bohemians), "monstres" (monsters), "pourritures" (rot), "manipulateurs" (manipulators), "parasites" (parasites), "salauds" (bastards), "socialistes" (socialists), "macronards" (Macron followers), "spikés" (spiked),

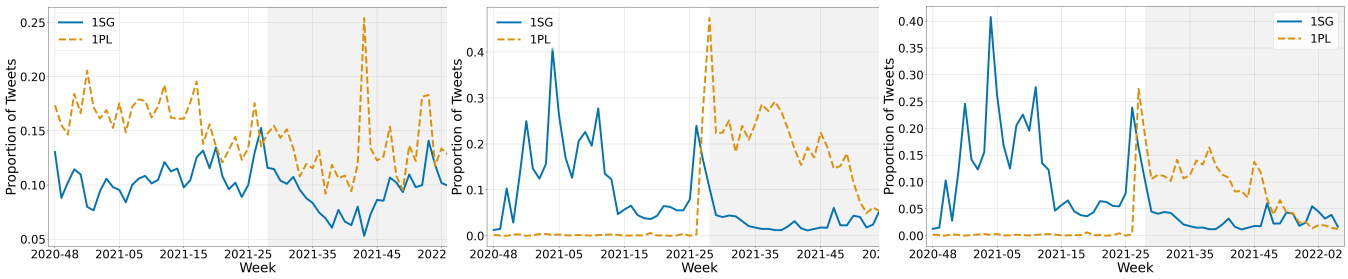


Figure 11: Pronoun use (A) excluding hashtags, (B) including only hashtags, and (C) including only the hashtags that were *not* used for data collection; with the period following Macron's speech shaded grey. The collective shift from first-person singular (1SG) pronouns to first-person plural (1PL) pronouns happens only when including hashtags in the analysis.

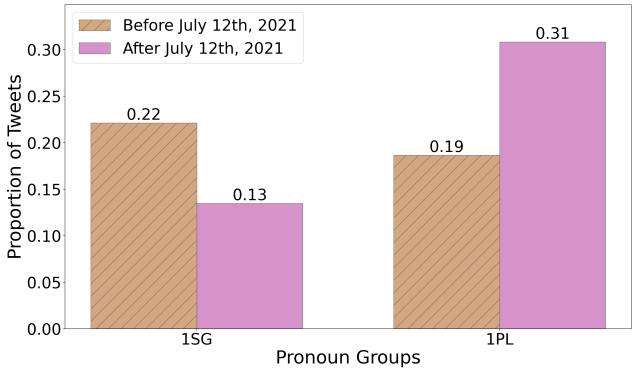


Figure 12: Pronoun use before versus after Macron's speech, considering two groups of pronouns: first-person singular (1SG) and first-person plural (1PL) pronouns. Before Macron's speech, users mostly used 1SG pronouns while they used more 1PL pronouns after.

"dosés" (dosed), "vaxxinés" (vaccinated), "labos" (labs), "journaleux" (journalists), "menteurs" (liars), "khmers" (Khmer), "hypnotisés" (hypnotized), "pleutres" (cowards), "toutous" (puppets), "veaux" (calves), "obéissants" (obedient), "ayatollahs" (ayatollahs), "godillots" (bootlickers), "sbires" (henchmen), "fascistes" (fascists), "toubibs" (doctors), "laboratoires" (labs), "guignols" (clowns), "voyous" (thugs), "talibans" (Talibans), "endoctrinés" (indoctrinated), "journalopes" (journalists), "lobbies" (lobbyists), "hystériques" (hysterical), "assassins" (assassins), "minables" (pathetic), "sorciers" (sorcerers), "apeurés" (fearful), "traitres" (traitors), "enfoirés" (bastards), "kapos" (kapos), "zélés" (zealous), "dingues" (crazy), "endormis" (sleepy), "cocus" (cuckolded), "dégénérés" (degenerates), "racailles" (scum), "migrants" (migrants), "incompétents" (incompetents), "nuls" (useless), "acolytes" (acolytes), "employeurs" (employers), "complices" (accomplices).

Incoming and Established Members

To compare different user groups, we complement our main analysis by reporting additional detail on cosine similarity values and their distribution. Table 5 presents the means, standard deviations, and bootstrap confidence intervals.

Sensitivity Analyses

Outgroup Labels Threshold

In the main analysis, we selected outgroup labels using a 75% threshold, meaning a word is considered an outgroup label if it occurs as such in at least 15 out of 20 tweets in both human annotations and GPT output. This resulted in a total of 94 outgroup labels. We tested this threshold by running the analysis on the prevalence of outgroup labels in different categories of authors with different thresholds. Each threshold reduced the number of outgroup labels retained: 16 (80 outgroup labels), 17 (69 outgroup labels), 18 (58 outgroup labels), 19 (40 outgroup labels), and 20 (13 outgroup labels). Figures 14, 15, 16, 17, 18 show the results for each threshold. While the absolute values change slightly, the overall dynamics remain consistent: Incoming-Transient users consistently exhibit the lowest proportion of outgroup labels, Established-Core users the highest, and Established-Irregular and Incoming-Persistent users show similar behavior across thresholds.

Cosine Similarity between Categories of Users

We tested the robustness of our author groupings to variations in the thresholds used to define Established-Core users. Varying the cutoff for "top 10% most active authors" to 15% or 20% resulted in negligible changes (±1 author) and did not affect the overall results. Adjusting the minimum activity duration threshold for Established-Core users (from $\geq$10 months to $\geq$8 or $\geq$6 months) changed the group size substantially (427 $\rightarrow$ 817 $\rightarrow$ 1555 Established-Core users). Figures 19 and 20 show, respectively, cosine similarity when varying the number of minimum active months to 8 and 6. As in the main analyses, Established-Core users remain the most self-similar group, Incoming-Transient users the least, while Established-Irregular and Incoming-Persistent users display a similar behavioral pattern.

Overview Figure

Figure 21 provides an overview of the research design. The diagram outlines the main steps of the study, beginning

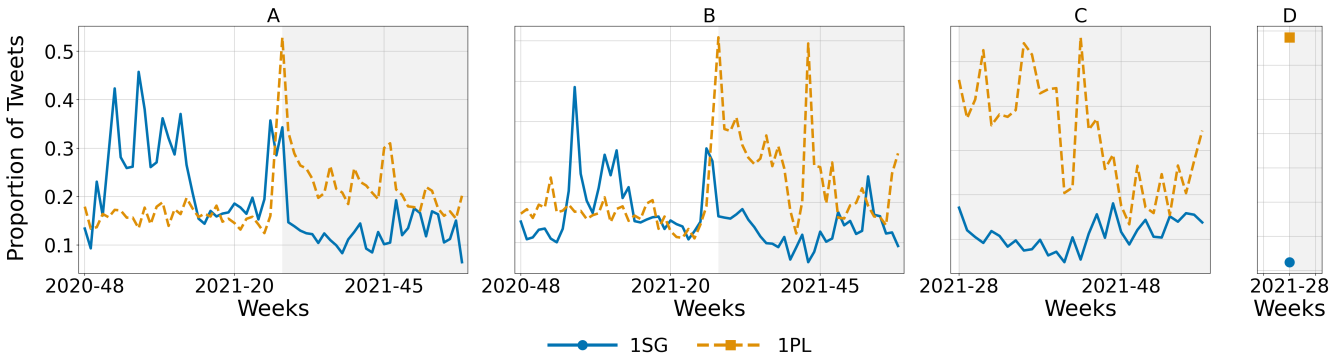


Figure 13: Pronoun use in tweets of (A) Established-Core, (B) Established-Irregular, (C) Incoming-Persistent, and (D) Incoming-Transient users; with the period following Macron's speech shaded grey. Established users use more first-person singular (1SG) pronouns before Macron's speech while all groups use more first-person plural (1PL) pronouns from the week of his speech.

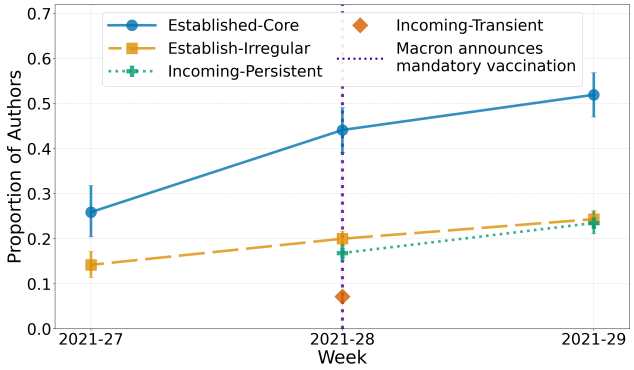


Figure 14: Proportion of outgroup label use by authors with a threshold of 16 out of 20 tweets

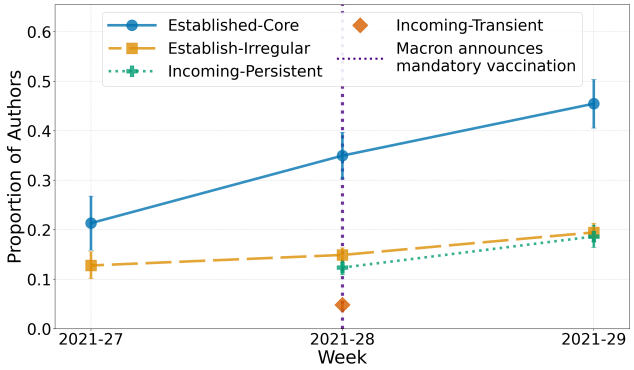


Figure 16: Proportion of outgroup label use by authors with a threshold of 18 out of 20 tweets

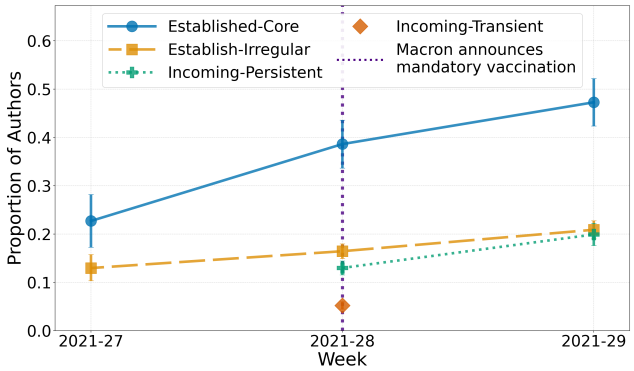


Figure 15: Proportion of outgroup label use by authors with a threshold of 17 out of 20 tweets

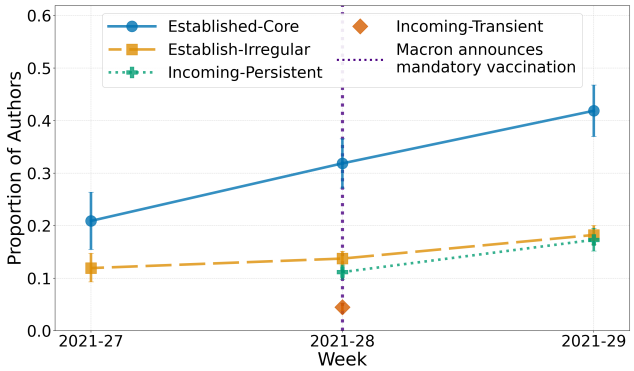


Figure 17: Proportion of outgroup label use by authors with a threshold of 19 out of 20 tweets

Table 5: Cosine similarity results with standard deviation and bootstrap confidence intervals.

Group1	Group2	Mean	SD	CI95_low	CI95_high	N	Week
Established Core	Established Core	0.509	0.144	0.508	0.510	48180	2021-27
Established Core	Established Irregular	0.436	0.142	0.436	0.437	125400	2021-27
Established Irregular	Established Irregular	0.379	0.138	0.378	0.379	324328	2021-27
Established Core	Established Core	0.595	0.160	0.594	0.596	146304	2021-28
Established Core	Established Irregular	0.504	0.170	0.504	0.504	897752	2021-28
Established Core	Incoming Persistent	0.486	0.164	0.486	0.487	792044	2021-28
Established Core	Incoming Transient	0.420	0.154	0.420	0.420	1271943	2021-28
Established Irregular	Established Irregular	0.438	0.177	0.438	0.438	5490132	2021-28
Established Irregular	Incoming Persistent	0.419	0.171	0.419	0.419	4847392	2021-28
Established Irregular	Incoming Transient	0.368	0.166	0.368	0.368	7784424	2021-28
Incoming Persistent	Incoming Persistent	0.404	0.166	0.404	0.404	4271564	2021-28
Incoming Persistent	Incoming Transient	0.352	0.161	0.352	0.352	6867828	2021-28
Incoming Transient	Incoming Transient	0.312	0.158	0.312	0.312	10997400	2021-28
Established Core	Established Core	0.636	0.151	0.635	0.637	148608	2021-29
Established Core	Established Irregular	0.547	0.161	0.546	0.547	707538	2021-29
Established Core	Incoming Persistent	0.539	0.165	0.539	0.540	455094	2021-29
Established Irregular	Established Irregular	0.473	0.164	0.473	0.474	3357934	2021-29
Established Irregular	Incoming Persistent	0.468	0.167	0.468	0.468	2161107	2021-29
Incoming Persistent	Incoming Persistent	0.464	0.170	0.463	0.464	1388770	2021-29

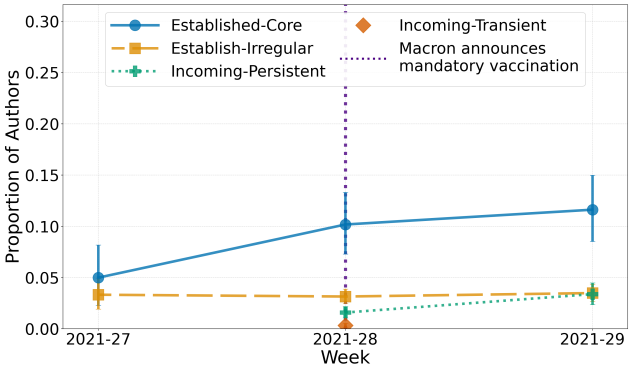
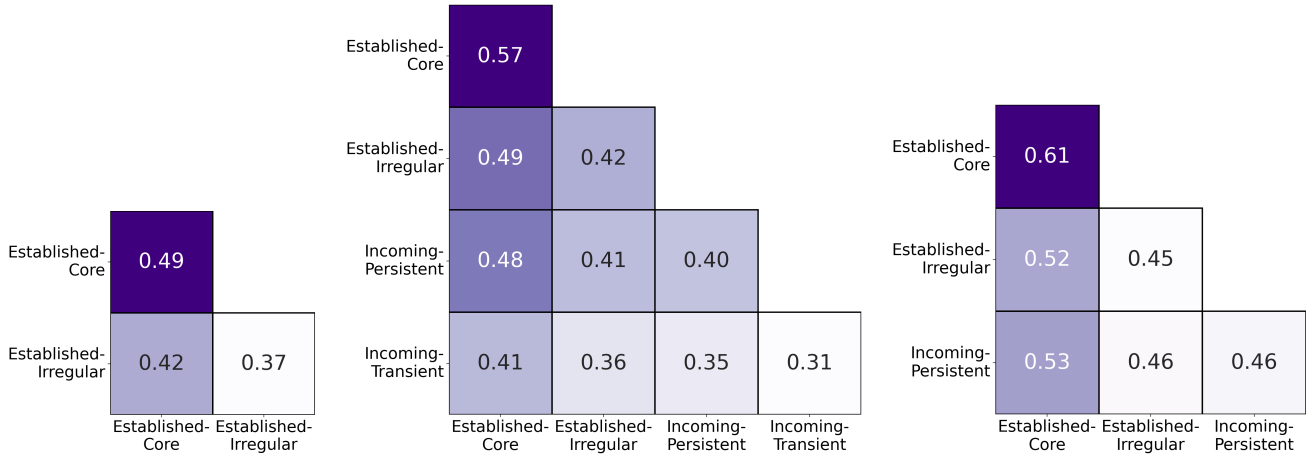


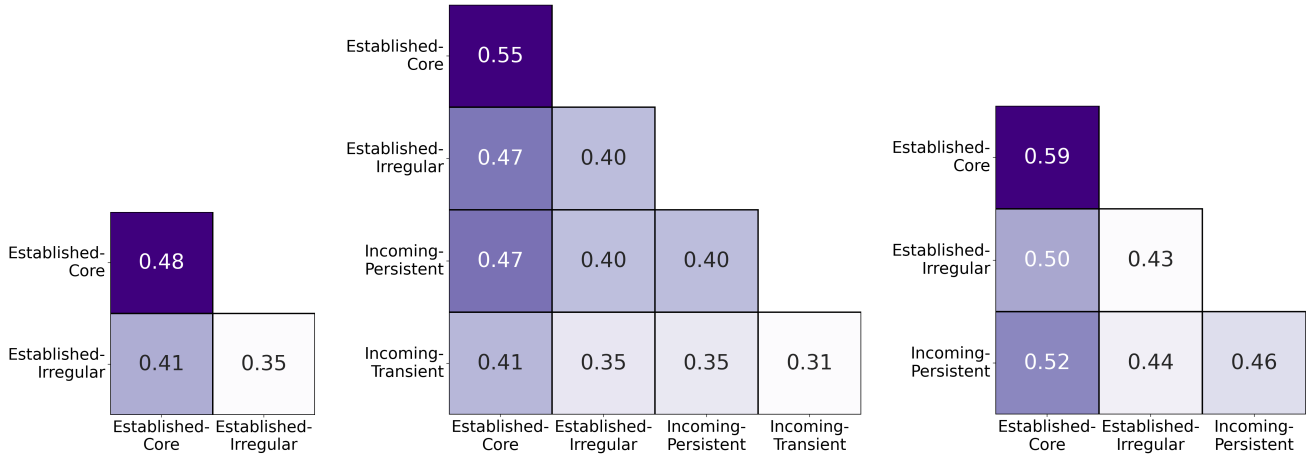
Figure 18: Proportion of outgroup label use by authors with a threshold of 20 out of 20 tweets

with data collection, followed by the three main analyses of pronouns, outgroup labels, and categorization of users. This overview clarifies the structure of the paper and each methodological component.



A - Week 2021-27 **B - Week 2021-28** **C - Week 2021-29**

Figure 19: Cosine similarity between Established-Core, Established-Irregular, Incoming-Persistent, and Incoming-Transient groups, when varying the threshold for Established-Core users from $\geq$10 to $\geq$8 active months. We focused on the week before Macron's speech (A - 2021-27), the week of his speech (B - 2021-28), and the week after (C - 2021-29).



A - Week 2021-27 **B - Week 2021-28** **C - Week 2021-29**

Figure 20: Cosine similarity between Established-Core, Established-Irregular, Incoming-Persistent, and Incoming-Transient groups, when varying the threshold for Established-Core users from $\geq$10 to $\geq$6 active months. We focused on the week before Macron's speech (A - 2021-27), the week of his speech (B - 2021-28), and the week after (C - 2021-29).

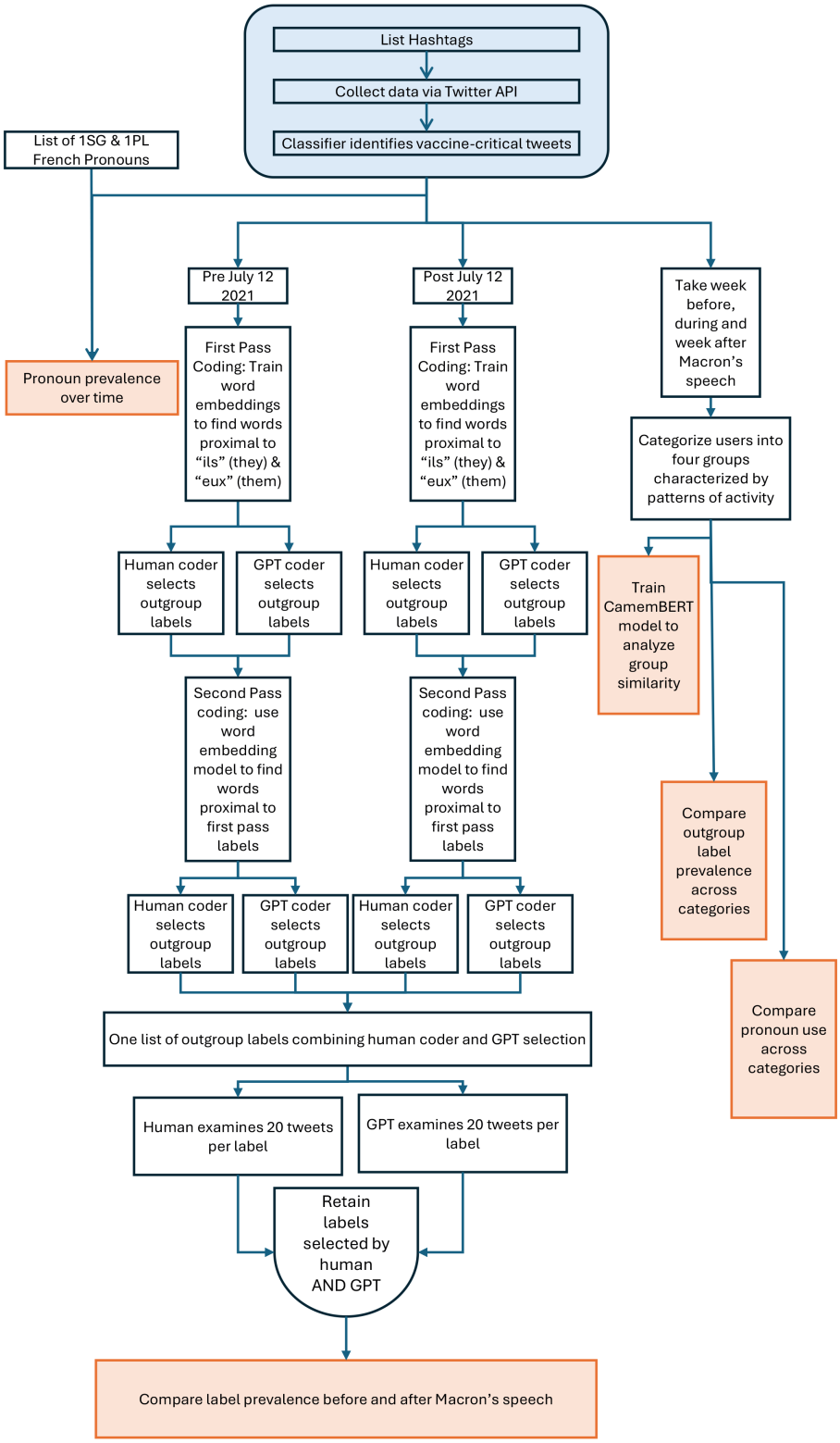


Figure 21: Overview of the research design.