

Emotional and Informational Trajectories of Immigrants: A Longitudinal Study of Reddit Communities

Anonymous submission

Abstract

Migration is often studied through surveys, administrative records, or clinical data, yet these sources offer limited visibility into how settlement unfolds in everyday life before and after arrival. We introduce a longitudinal analysis of self-reported immigrants to Canada on Reddit, treating migration as a temporally anchored life-course transition rather than a single event. Using keyword-based retrieval and an LLM-assisted temporal inference pipeline, we reconstruct pre-arrival, arrival-year, and post-arrival timelines for 9,834 immigrant Reddit users who arrived between 2014 and 2024, comprising 4,861,954 posts and comments.

Combining LIWC-based psycholinguistic analysis with embedding-based topic clustering, we find that overt affective change around arrival is modest: contrary to expectations from migration mental-health literature, negative emotion barely shifts, while positive emotion declines reliably from arrival to post-arrival with a small effect. The stronger linguistic signal is a reorientation of self-positioning: after arrival, newcomers write less about themselves and more about families, communities, and external institutions. Topical attention, in contrast, shifts sharply and structurally. Around arrival, leisure, technology, and media-oriented discussion give way to settlement geography, immigration procedures, jobs, and housing; after arrival, acute procedural concerns recede while infrastructural ones such as cost of living, transport, and finance remain elevated.

Together, these findings show that migration reshapes Reddit discourse less through dramatic emotional change than through a durable reallocation of attention toward the relational, institutional, and material work of settlement. Methodologically, our study offers a scalable framework for temporally anchoring social media data around life-course events; substantively, it reveals when newcomer information needs emerge, with implications for the timing of settlement support.

Introduction

Immigration in the global north countries including Canada is both a macro-level policy instrument and a deeply personal trajectory that unfolds through uncertainty, anticipation, and adaptation. In Canada, the annual immigration targets in the Supplementary Immigration Levels Plan project approximately 395,000 permanent residents in 2025, 380,000 in 2026, and 365,000 in 2027 (Immigration and Canada 2025). These numbers reflect Canada's reliance on

migration for demographic renewal and economic growth, underpinned by policy frameworks that explicitly center multiculturalism and diversity as national values (Can 1988; of Canada 1971). As a result, newcomers to Canada bring a wide range of professional backgrounds, linguistic repertoires, and migration histories.

Yet, arrival in a “welcoming” system does not guarantee an easy transition. Newcomers must simultaneously navigate employment, housing, credential recognition, language learning, and evolving support networks. These pressures are often intensified by precarious legal status, information asymmetries, and under-resourced settlement infrastructures. Early settlement experiences can therefore involve isolation, uncertainty, and emotional strain, with downstream consequences for economic participation, social integration, and mental wellbeing (Monteiro 2022). Understanding how immigrants cope with these challenges, and how their emotional states evolve from pre-arrival planning to everyday life in Canada, is central to migration scholarship and agendas aimed at supporting vulnerable populations online.

Traditional data sources for studying immigrant wellbeing, including surveys, administrative records, or clinical health records, offer important but limited coverage. Surveys capture narrow time slices and are costly to repeat; administrative and clinical data often miss those who are under-insured, reluctant to seek care, or navigating precarious status. In contrast, social media platforms provide ongoing, time-stamped traces of how people ask questions, seek support, vent frustrations, and share milestones in everyday language. Prior work shows that social media language can encode stress, depression, coping behaviours, and support-seeking at population scale (Saha et al. 2022; Yuan et al. 2023; Kamarudin et al. 2020; Sabri et al. 2024; Garg et al. 2023), suggesting that emotional trajectories around major life events may be observable in digital traces.

We focus on Reddit as our case platform. Reddit is pseudonymous, discussion-oriented, and organized around topic-based communities rather than offline identity, which lowers the friction of candid disclosure. In immigration-focused subreddits, users discuss visa timelines, housing, employment, health insurance, and loneliness, and often substitute Reddit for formal newcomer support systems (Mittal et al. 2023; Monteiro 2024). A newcomer can, for example, ask late at night whether their work-permit de-

lay is typical, how to interpret a tax slip, or how others coped with homesickness. These conversations reveal not only what newcomers know, but what they worry about and how their priorities shift over time.

However, most existing computational work either examines discourse *about* immigrants (e.g., anti-immigrant framing and incivility (Stravato Emes 2024; Ahmed et al. 2024)) or take cross-sectional snapshots of immigrant communities (Phan 2023; Mittal et al. 2023). Analyses that follow migrants longitudinally, from planning migration to settling in the destination country, are rare, in part because it is difficult to infer temporal anchors such as “year of arrival” from unstructured text. As a result, we know relatively little about how migrants’ emotional expression and topical priorities evolve before and after arrival, or how these trajectories vary across policy and economic contexts.

Our Approach and Research Questions

In this work, we treat migration to Canada as a natural life-course intervention and ask how does online language shift around this transition? We assemble a large Reddit corpus using high precision keywords as entry points, and use an LLM-based pipeline¹ to (i) identify users who self-report as immigrants to Canada, and (ii) infer their year of arrival by resolving relative temporal expressions (e.g., “moved last year”) against post timestamps. **This enables us to reconstruct pre-arrival, arrival-year, and post-arrival timelines for 9,834 self-reported immigrant Reddit users, spanning 4,861,954 posts and comments.**

On top of these timelines, we apply LIWC-based psycholinguistic markers to operationalise emotional expression, and an embedding-based clustering pipeline to surface high-level settlement topics. Analytically, we adopt a *quasi-experimental* mindset where for each user, we compare linguistic markers within-person across periods, using pre-arrival and arrival-year activity as baselines and post-arrival activity as outcome. Our analysis focuses on two main research questions (RQs):

- **RQ1:** How does immigration influence emotional and psychosocial expression as reported on Reddit before, during, and after arrival in the host country?
- **RQ2:** How does immigration reshape topical priorities in Reddit discourse, and how do these themes evolve across the arrival boundary and over calendar time?

Beyond offering a large-scale description of immigrant discourse, we are interested in what *social media data* uniquely reveals about settlement as it unfolds. Reddit posts are not structured survey responses or clinical notes; they are self-initiated traces of what people choose to ask, vent about, or celebrate in public. Treating these traces as longitudinal signals allows us to observe how the same individuals move from anticipation to bureaucratic negotiation and,

eventually, to everyday life. In doing so, we aim to demonstrate how naturally occurring web data can be mobilised in the service of migrant well-being and integration.

Contributions

Building on the above design, this paper makes three contributions:

1. **An expanded temporally anchored immigrant Reddit corpus.** We construct a longitudinal dataset of self-reported immigrants to Canada with inferred arrival years, retaining authors with validated immigrant status and temporal anchors. This enables pre-arrival, arrival-year, and post-arrival comparisons across emotional, psychosocial, and topical dimensions of migrant discourse.
2. **A mixed-methods analysis of migration-linked psychosocial expression.** Using LIWC-based psycholinguistic markers, we show that migration is associated less with dramatic changes in overt emotional language than with subtler shifts in self-positioning and social orientation. Post-arrival discourse becomes less individually self-focused and more outward-facing, relational, and collectively referenced.
3. **Evidence of phased topic shifts and design implications.** Through semantic topic clustering, we show how migrants’ concerns shift from preparation, technology, leisure, and everyday routines before arrival, to immigration processing, employment, finance, housing, settlement geography, and transport during arrival, and toward diaspora/community, healthcare, mobility, politics, cost of living, and everyday embedded life after settlement. We discuss how these patterns can inform timely newcomer support and online settlement infrastructures.

The rest of the paper proceeds as follows. Section 2 situates our work within research on immigration, mental health, and social media. Section 3 details our data collection, immigrant identification, arrival-year inference, emotional measurement, and topic modeling procedures. Section 4 presents our findings for RQs 1 and 2. Section 5 discusses our findings in respect to broader critical literature, along with limitations and future directions. Section 6 concludes the paper.

Related Work

We briefly review three strands of work that inform our study: (1) mental health analysis using social media, (2) Reddit and immigrant communities, and (3) research on migration, integration, and web-mediated immigration discourse, and existing gaps in the literature.

Mental Health Analysis Using Social Media: A substantial body of research demonstrates that social media language can encode markers of stress, depression, anxiety, and broader mental health concerns. Reddit, has been widely used because its topic-specific communities and semi-anonymous structure facilitate candid self-disclosure and large-scale analysis of psychological distress and well-being (Saha et al. 2022; Yuan et al. 2023; Kamarudin et al.

¹All LLM prompts and settings are provided in Appendix B. GitHub anonymized repository available at <https://anonymous.4open.science/r/immigration-trajectory-reddit-114E/>

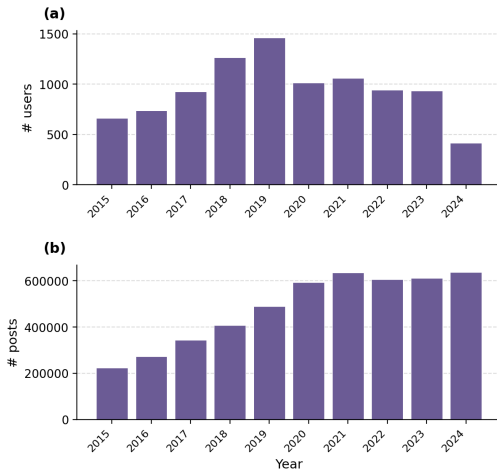


Figure 1: Distribution of data; (a) Author arrival years, (b) number of posts/comments by year.

2020). Recent advances incorporate neural models, specialized mental-health datasets, and LLM-driven approaches for more context-aware prediction, including corpora such as LoST for low self-esteem (Garg et al. 2023), multi-task models of burnout discourse (Sabri et al. 2024), and LLM-based frameworks like MentaLLaMA (Yang et al. 2024).

Within this broader space, Saha et al.’s quasi-experimental analysis of Twitter during the early COVID-19 pandemic demonstrates how social media can be used to quantify temporal changes in mental health-related language around a major exogenous event (Saha et al. 2020). They compare over 60 million tweets from March–May 2020 with a pre-pandemic control period, using classifiers trained on Reddit mental-health communities to estimate shifts in anxiety, depression, stress, and suicidal ideation, and SAGE to characterise COVID-specific lexical changes. Subsequent work extends this approach to college students during COVID-19 (Saha, Kotakonda, and De Choudhury 2025), leveraging college subreddits to examine mental health trajectories in a demographically and contextually bounded population.

Such work in the literature shows that (i) social media traces can serve as useful proxies for tracking temporal changes in emotional expression at scale, and (ii) quasi-experimental comparisons around exogenous events can sharpen causal interpretations. Our work adopts this temporal and quasi-experimental orientation, but centers on a particular kind of event where the focus is on individually experienced migration to a new country, rather than a shared global crisis.

Reddit for Migrant Communities: A parallel line of work examines immigration and migration through online discourse. Emes investigates how immigrants and immigration are framed in online conversations mirrored from Singapore news comment sections onto Reddit, identifying high-level interpretive frames and training a supervised classifier to quantify their prevalence (Stravato Emes 2024). Such work demonstrates that immigration discourse is structured by recurring narrative patterns (e.g., threat, economic bur-

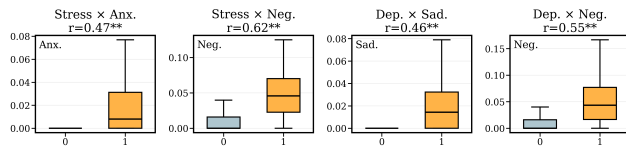
den, cultural difference) that can be captured at scale. Related multi-method research shows how social media can amplify or reshape anti-immigrant prejudice and hostility (Ahmed et al. 2024).

Other work examines Reddit as a venue for sustained immigration discussion. Phan’s work clusters a decade of Reddit posts about immigration programs to study sentiment and thematic structure (Phan 2023). Mittal et al. focus directly on immigrant communities, showing that immigration-related subreddits contain distinct mental health signatures compared to general communities and operate as informal support systems for migrants (Mittal et al. 2023). These studies highlight Reddit as a key arena where immigration is debated and emotionally processed, but most analyses remain either aggregate (subreddit-level) or cross-sectional.

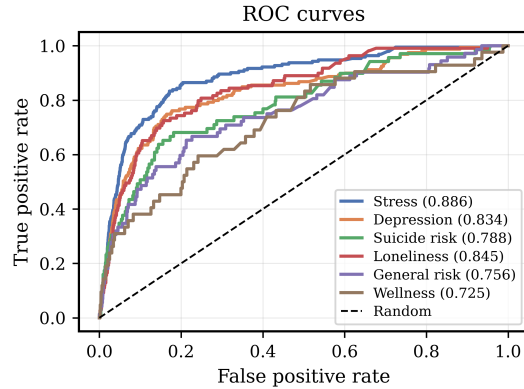
In the Canadian context, Monteiro’s work shows that newcomers use Reddit extensively to seek information about settlement processes and services, often treating the platform as an extension of newcomer support infrastructure (Monteiro 2024). Other work links social media use to employment outcomes among newcomers (Monteiro 2022) or examines representations of migrants in news shared on Reddit (Coppin 2020). These studies show that digital platforms are both sites of public discourse about immigration and practical infrastructures for settlement.

Web-Mediated Immigration Discourse: Beyond specific platforms, research on migration, integration, and health on social media and the broader web explores how online discourse captures trajectories of mental health, social support, and integration challenges. Nguyen et al. chart a decade of migration-related health and integration discourse, highlighting temporal trends in sentiment and topics (Nguyen et al. 2025). Analyses of digital nomads’ careers via Reddit illustrate how mobility decisions, career narratives, and identity work unfold in online conversations (Zikic, Zupic, and Černe 2025). Work on code-mixing in migration discourse examines how migrants’ linguistic strategies reflect identity negotiation and social positioning across platforms (Vitiugin et al. 2024).

Within the existing literature, **we identify a methodological gap**. While prior work (1) tracks changes in mental-health-related expression around major events using social media (Saha et al. 2020; Saha, Kotakonda, and De Choudhury 2025; Saha et al. 2022; Yuan et al. 2023; Yang et al. 2024), (2) documents immigration discourse and informal support systems on Reddit (Stravato Emes 2024; Phan 2023; Mittal et al. 2023; Ahmed et al. 2024; Monteiro 2024), and (3) highlights the role of digital platforms in integration processes (Monteiro 2022; Nguyen et al. 2025), we still lack longitudinal analyses that follow migrants from pre-arrival planning into post-arrival settlement. In particular, few studies infer individual-level migration timelines and link them to emotional and topical trajectories. Our work addresses this gap by reconstructing arrival timelines for self-disclosed immigrant Redditors, and by using these timelines to examine how emotional expression and everyday concerns evolve across the migration boundary.



(a) Rank-biserial correlations



(b) AUC scores

Figure 3: Panel (a) compares LIWC score distributions for MentaLLaMA-negative (0) and MentaLLaMA-positive (1) posts across four conceptually aligned constructs. Sample sizes are 95,121/192 for stress and 95,154/159 for depression.

82.5% of cases; 14.6% are judged as “off by 1” (often off by one year near the turn of the calendar), and 2.9% are clearly incorrect. For these 9,834 users, we obtained 4,861,954 posts and comments in total. Activity is high (mean = 563.8 posts/author; median = 269; min = 5; max = 4,887), providing sufficient temporal depth for within-person analysis. Figure 1(b) shows how the data is distributed across the different calendar years, and Figure 1(a) shows how arrival years for each author are distributed within the same time frame. We leave out data prior to 2015 since Reddit was not as heavily used, so data is more scarce and often more niche, and we also leave out 2025 data because we did not collect data for the entire year.

Measuring Psycholinguistic Emotional Expression

To estimate changes in emotional and psychosocial expression across the migration timeline, we rely on psycholinguistic features derived from the Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker 2010). Prior work has established LIWC as an interpretable and robust tool for tracking affective and cognitive signals in social media text (Tausczik and Pennebaker 2010; Saha, Kotakonda, and De Choudhury 2025). We intentionally eschew classifier-based symptom detection in favour of LIWC, prioritising transparency, temporal stability, and linguistic generalisability across the diverse sociocultural backgrounds represented in Canadian immigration discourse.

However, we do also recognize that LIWC is a dictionary-based approach, and may not fully capture the cultural and contextual nuance that SOTA transformer-based mod-

els might. To address this concern and in order to show that our adoption of LIWC is consistent with other transformer based models, we conducted an additional validation analysis using MentaLLaMA (Yang et al. 2024), a state-of-the-art transformer-based mental health classifier, designed to capture contextual emotional signals beyond keyword matching. We annotated a sample of 95,313 posts with both MentaLLaMA and LIWC, and our analysis show that across multiple constructs, posts identified by MentaLLaMA as reflecting stress, depression, loneliness, suicide risk, or wellness show strong and statistically significant associations with conceptually aligned LIWC dimensions. We report rank-biserial correlations because LIWC category scores are highly skewed and frequently zero-inflated at the post level, making rank-based effect sizes more appropriate than Pearson-derived alternatives. Rank-biserial r measures the extent to which posts identified by MentaLLaMA tend to receive higher LIWC scores on conceptually aligned dimensions (Figure 3.a). Posts flagged for stress were associated with elevated LIWC negative emotion (rank-biserial $r=0.62$) and anxiety ($r=0.47$), while depression-related posts aligned strongly with LIWC negative emotion ($r=0.55$) and sadness ($r=0.46$). In addition, LIWC-derived features predicted MentaLLaMA labels with AUC values ranging from 0.73 to 0.89 across constructs, substantially above chance (Figure 3.b). These findings suggest that the broad temporal patterns reported in our study are not solely artifacts of dictionary matching, but are also reflected in a contextual transformer-based model sensitive to richer linguistic cues. We therefore view LIWC as a scalable and interpretable approximation whose major signals show meaningful convergence with more context-aware methods.

As such, each Reddit post and comment is processed through LIWC to obtain category scores, including *Affect*, *Negative Emotion*, *Anxiety*, *Anger*, *Sadness*, *Cognitive Processes*, and pronoun-related categories. Scores reflect the proportion of words in the text that belong to a given LIWC category, yielding continuous measures of emotional expression rather than binary labels. Because LIWC is dictionary-anchored rather than trained on a specific dataset, it is less susceptible to temporal drift in Reddit vocabulary and avoids strong assumptions about clinical status. Throughout, we treat LIWC outputs as markers of emotional language, not diagnoses.

Topic Modeling and Semantic Clustering

To complement our emotional analysis with a thematic view of migrant concerns, we implement a semantic clustering pipeline over the full set of Reddit posts/comments. We first encode each text using the all-MiniLM-L6-v2 sentence transformer. This model offers a strong trade-off between semantic accuracy and computational efficiency, making it suitable for large-scale social media corpora (Wang et al. 2020). Compared to bag-of-words or TF-IDF representations, dense sentence embeddings better capture paraphrases and contextual nuance in short, informal text (Reimers and Gurevych 2019).

Since high-dimensional embeddings can hinder clustering, we apply Principal Component Analysis (PCA) to re-

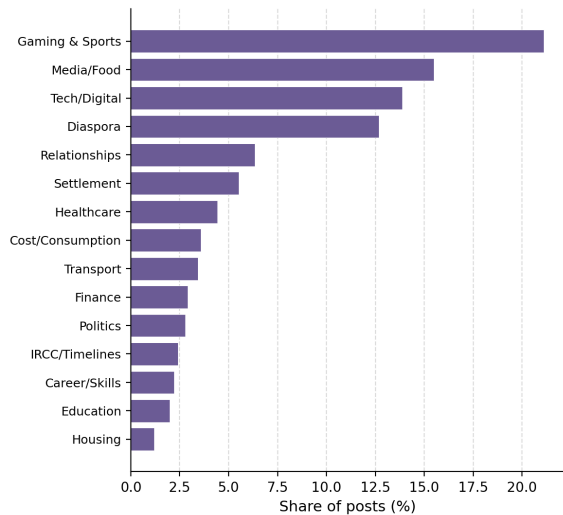


Figure 4: Distribution of Reddit posts across topic categories.

duce noise and improve cluster separability prior to clustering (Ding and He 2004). We then cluster the resulting vectors using KMeans (MacQueen 1967), chosen for three reasons: (i) it scales linearly to over 100k embeddings; (ii) it produces discrete, non-overlapping clusters amenable to quantitative comparison; and (iii) it aligns well with PCA-projected spaces.

We perform a grid search over candidate numbers of clusters k and evaluate each setting using *silhouette* scores (Rousseeuw 1987). While $k = 100$ yields a slightly higher score, it produces overly fragmented clusters that are difficult to interpret. We therefore select $k = 40$ as a balance between model fit and interpretability. We manually inspect cluster centroids and representative posts, merging semantically overlapping clusters and discarding noise-dominated groups (e.g., link-only replies or generic chatter). This human-in-the-loop consolidation is consistent with recent embedding-based topic modeling approaches that emphasize semantic coherence and interpretability (Grootendorst 2022). The final topic clusters and their distributions are shown in Figure 4.

In total 1,363,959 out of 4,861,954 posts and comments (approximately 28%) are mapped to one of these topical categories. We use this mapping to compare topic distributions across pre-arrival, arrival-year, and post-arrival periods.

Findings

We now present our findings, organised by research question. We first examine how emotional expression changes around arrival (RQ1), then analyse shifts in topical priorities across the migration timeline and over calendar time (RQ2).

RQ1: Emotional Trajectories Around Arrival

To examine how immigrants’ emotional, cognitive, and social expression changes around migration, we analyze within-user shifts in LIWC-derived markers across three migration windows (pre-arrival, arrival-year, and post-

arrival). Table 1 reports percentage-point changes alongside Bonferroni-corrected Wilcoxon signed-rank tests and Cohen’s effect sizes. Because the cohort is large enough that small differences can register as statistically reliable, we interpret results by triangulating significance, absolute change, and effect size.

A major observation that frames this section is that the strongest within-user signal in our data is not an affective one. Categories conventionally tied to emotional valence, including positive and negative emotion, anxiety, sadness, and anger, move modestly across most comparisons, and where they reach significance their effect sizes remain below conventional thresholds. Pronoun categories, in contrast, show a coherent and comparatively large reorganization, with first-person singular declining sharply post-arrival while collective *we* and external *they* rise across every phase comparison. In psycholinguistic work, pronouns are treated as markers of psychological orientation rather than topic, with first-person singular indexing self-focus and inwardness and *we*-words and *they*-words indexing group identification and externalization (Tausczik and Pennebaker 2010). Through this lens, the migration trajectory in our data is one in which speakers reposition themselves in language, not one in which they grow more or less emotionally expressive online.

Pre-arrival to arrival-year. Moving into the arrival year produces almost no affective change. Positive emotion drifts down slightly without reaching significance, and negative-affect categories remain effectively flat. Anxiety registers a small but statistically significant increase that is consistent with the anticipatory uncertainty long documented in qualitative work on the pre-migration period (Ager and Strang 2008). The more interpretable changes in this window are relational. Collective *we* and external *they* both rise with small but reliable effects, family terms increase modestly, and social-process and friend-related language move only marginally. What this profile resembles is an anticipatory reorientation in which speakers begin to frame their lives through broader social units, including the family they are moving with, the diasporic community they are about to join, and the people they are leaving behind, before any clear affective shift takes place. Migration scholarship has long argued that the move is socially imagined and rehearsed well before it is emotionally lived (Carling and Collins 2020), and our data echo that idea at the level of language. The arrival year in our corpus is therefore better understood as the moment at which the unit of self-reference begins to widen, rather than as a moment of emotional uplift or strain.

Arrival-year to post-arrival. The post-arrival phase is where the affective signal in our data is concentrated, and even here it is modest. Positive emotion declines reliably and with a non-trivial effect, representing the largest substantive affective change in the table. Negative-affect categories shift only slightly, and while several reach significance their effect sizes stay small, so we treat them as descriptive rather than primary. The dominant change in this transition is pronominal. First-person singular falls sharply and produces one of the largest within-user effects in the table, while collective and external pronouns continue to rise.

Table 1: Percentage-point (pp) changes in LIWC-derived mental health markers across migration phases. Δ denotes absolute percentage-point differences between phases; bars are scaled to a number close to the maximum absolute pp change (0.3) observed in the data (0.5) for better readability. Green bars indicate positive changes and red bars indicate negative changes. d denotes Cohen’s d effect size; stars indicate statistical significance from Bonferroni-corrected Wilcoxon signed-rank tests (* $0.01 < p < 0.05$, ** $0.001 < p < 0.01$, *** $p < 0.001$).

LIWC	Pre → Arrival ($N=3,932$)		Arrival → Post ($N=5,143$)		Pre → Post ($N=3,674$)	
	Δ (pp)	d	Δ (pp)	d	Δ (pp)	d
posemo	-0.13	-0.029	-0.36	-0.090***	-0.42	-0.100***
negemo	-0.06	-0.034	0.04	0.023*	-0.02	-0.011
anx	0.01	0.037**	-0.01	-0.019***	0.01	0.019***
sad	-0.02	-0.026	0.01	0.007***	0.00	0.001***
anger	-0.04	-0.034	0.02	0.016***	-0.04	-0.032
family	0.03	0.037*	-0.02	-0.024***	0.02	0.032***
friend	-0.03	-0.024	-0.06	-0.055	-0.09	-0.113***
social	0.07	0.018	-0.20	-0.050	-0.09	-0.024
insight	-0.01	-0.007	-0.02	-0.016	-0.01	-0.004
cause	-0.03	-0.014	0.04	0.040***	0.02	0.013**
discrep	0.03	0.027	-0.03	-0.032	-0.01	-0.012
tentat	0.03	0.019	-0.04	-0.022	-0.02	-0.010
certain	0.02	0.019	0.04	0.022***	0.06	0.030***
i	0.05	0.020	-0.34	-0.155***	-0.23	-0.098***
we	0.04	0.061***	0.01	0.017***	0.04	0.071***
they	0.06	0.055***	0.03	0.049***	0.10	0.095***
you	-0.01	-0.005	-0.06	-0.024	-0.06	-0.023

Social-process and friend-related language drift downward without reaching significance. What this profile suggests is not an intensification of negative feeling but a redistribution of attention away from the self. Newcomers in our cohort continue to write at high volume, yet they write about themselves less as individuals. Such a redirection is consistent with sociological accounts of early settlement in which the practical demands of institutional navigation, employment search, and the rebuilding of social ties pull cognitive and discursive attention outward (Ager and Strang 2008; Moynihan, Herd, and Harvey 2015). The pronoun signature we observe is what one would expect when everyday life is dominated by others and by external systems rather than by introspection.

Pre-arrival to post-arrival. The longer pre-arrival to post-arrival comparison confirms that the cumulative linguistic effect of migration is a reorientation rather than a deterioration. Positive emotion declines reliably with a modest effect size, while negative emotion essentially does not move. The clearest long-term changes are relational. First-person singular declines and collective *we* and external *they* both rise, while the single largest effect in the long-term column belongs to a steady decline in friend-related language across the migration window, accompanied by a modest increase in family-related language. This combination is substantively interesting because it points to a long-term recomposition of how immigrants’ relational world is referenced in writing. Over the course of migration, the typical author in our cohort becomes less likely to write in the first-person singular and less likely to invoke friends, while becoming more likely to invoke collective groups, external others, and family. We do not read this as a claim that immigrants lose

friends or grow closer to kin in real life. What the data show is a shift in how the relational world is articulated, which is consistent with sociological work on how migration reorganizes everyday networks and pulls family-of-origin and ethnic-community ties back into discursive salience (Anthias 2013; Leurs and Smets 2018).

Summary. Three observations consolidate this section before we turn to topical evidence in RQ2. First, affective change around migration in our cohort is modest. Positive emotion declines reliably after arrival but with a small effect, while negative emotion, sadness, and anger remain essentially unchanged across every phase comparison. The linguistic signature of migration in this population is therefore not one of intensified negative feeling. Second, the most consistent and structurally coherent change associated with migration in our data is a recomposition of relational and pronoun categories, with first-person singular declining and collective *we* and external *they* rising across every phase comparison, alongside a steady decline in friend-related language and a modest increase in family-related language. Third, this relational shift begins before affective change does, with collective and external pronouns rising during the arrival year while affective markers are still essentially flat, and continuing to rise into the post-arrival period when positive emotion declines. The overall picture is one in which migration reshapes the speaking subject in immigrants’ Reddit discourse. The dominant signal is not that immigrants come to express more or less emotion online, but that they come to narrate their lives less as individuals and more in relation to families, groups, and others. This recomposition previews the second part of our analysis, in which we examine whether the topics that immigrants attend to undergo a

Table 2: Cleaned representative keywords summarizing each topic. Generic fillers and duplicates were removed to highlight meaningful semantic indicators.

Topic	Representative Keywords
IRCC	waiting, processing time, update, application, apply, PR, visa, permit, passport, citizenship, work permit
Career / Skills	job, experience, company, resume, career, interview, working, employer, skills, profession
Education	course, exam, program, class, student, university, study, semester, school, test
Housing	rent, apartment, landlord, lease, condo, property, housing, rental, room, home
Cost / Living	price, cost, buy, money, store, shipping, usd, cad, amazon, expenses
Finance	account, credit, bank, refund, money, tax, tfsa, rrsp, mortgage, savings, income, investment, funds
Politics	economy, government, market, companies, business, policy
Settlement / Cities	city, area, neighborhood, park, downtown, Toronto, Vancouver, Montreal, Ontario, Quebec
Healthcare	doctor, pain, symptoms, weight, help, health concern
Relationships	relationship, friends, girlfriend, women, love, dating, friendship, social life
Transport	car, driving, parking, traffic, driver, road, lane, insurance, vehicle, bike, Toyota
Gaming / Sports	game, players, season, play, league, match, pc, gaming, upgrade, deck
Media / Food	food, recipe, chicken, pizza, taste, eating, movie, watch, song, series, episode, book
Tech / Digital	issue, fix, problem, update, app, screen, using, google, mobile, data, android
Diaspora	india, indian, china, chinese, america, american, israel, culture, language markers (na, ang, lang, ako, po, hindi), french articles (le, la, les), spanish (el, para, que), portuguese (não, mais, para)

comparable phased reorganization.

RQ2: Settlement Concerns and Information Needs

To characterize how immigrants’ substantive concerns change across migration stages, we apply topic modeling to the corpus and manually consolidate the resulting clusters into interpretable themes. Table 2 summarizes the final topic set and representative keywords, while Table 3 reports within-user percentage-point changes in topic prevalence across migration phases. We use paired Wilcoxon signed-rank tests with Bonferroni correction to assess whether phase-level shifts are statistically reliable, and report Cohen’s h to compare effect magnitudes across topics that differ substantially in base rate. Fine-grained changes are in Appendix C, while representative samples are provided in Appendix D.

A first observation that frames this section is that the topic-level signal is much sharper and more structurally coherent than the affective signal examined in RQ1. Several topic categories shift by multiple percentage points within the same individuals, with effect sizes an order of magnitude larger than those of the LIWC categories, and the directionality of the shifts forms a clear pattern. Almost every instrumental settlement topic, including settlement geography, immigration procedures, jobs, finance, healthcare, housing, education, transport, and cost of living, rises dur-

ing the migration window, while leisure and general-digital topics, including gaming, technology, and media-and-food, decline. This pattern places our analysis squarely in the tradition of work that treats settlement as a multi-domain process spanning institutional, economic, social, and material concerns (Ager and Strang 2008), and it is consistent with prior characterizations of Reddit as an informal settlement infrastructure through which newcomers manage these demands (Monteiro 2024; Mittal et al. 2023).

Pre-arrival to arrival-year. The pre-arrival to arrival transition is the most consequential of the three windows, producing a substantial reorganization of immigrants’ topical attention. Settlement-geography discussion shows the single largest increase, followed by immigration procedures, relationships, jobs, cost of living, healthcare, finance, education, housing, and transport. The breadth of this shift is striking. Effectively every domain conventionally associated with newcomer adaptation rises at the same time within the same authors, suggesting that arrival operates as a coordinated reallocation of online attention rather than a focal change in any one area. At the same time, gaming shows the largest single decline of any topic in the table, followed by technology and media-and-food. The picture that emerges is one in which general-interest and leisure discourse gives way to a dense cluster of settlement-specific concerns as the move becomes imminent and then takes place. This is consistent with literature on the informational demands of migration, which emphasizes that newcomers must rapidly absorb large volumes of bureaucratic, economic, and practical information around the point of entry, often in the absence of formal support (Monteiro 2024; Dekker et al. 2018). What our data show is that this informational labor is visible in the topical composition of an immigrant’s everyday online activity, and that its onset is sharp rather than gradual.

Arrival-year to post-arrival. The arrival to post-arrival transition is smaller in scale than the pre-arrival pivot, but it reveals a meaningful sorting of which concerns persist and which recede. Several procedural and acute-settlement topics decline, including immigration procedures, jobs, healthcare, education, and relationships. These are domains in which the most intense engagement is concentrated around the moment of entry, when applications, credential evaluations, initial job search, and the establishment of new social ties dominate everyday life. In parallel, other domains continue to rise, including transport, cost of living, settlement geography, finance, housing, and politics, alongside a partial rebound in media-and-food discussion. The contrast suggests a movement from an acute-settlement phase, in which bureaucratic and entry-related concerns dominate, to a more embedded phase, in which the structural costs of daily life in Canada become more salient. The post-arrival rise of cost of living, transport, and housing is particularly notable in the Canadian context, where affordability and urban mobility have become defining features of newcomer experience over the period covered by our data. The emergence of politics as a significant topic only in this transition is also interpretively important, since it suggests that civic and policy engagement develops once the more immediate practicalities of arrival

Table 3: Percentage-point (pp) changes in topic prevalence across migration phases. Δ denotes absolute percentage-point differences between phases; bars are scaled to the maximum absolute pp change observed (5.6). Green bars indicate positive changes and red bars indicate negative changes. $|h|$ denotes the absolute value of Cohen’s h effect size; stars indicate statistical significance from Bonferroni-corrected Wilcoxon signed-rank tests (* $0.01 < p < 0.05$, ** $0.001 < p < 0.01$, *** $p < 0.001$).

Topic	Pre → Arrival		Arrival → Post		Pre → Post	
	Δ pp	$ h $	Δ pp	$ h $	Δ pp	$ h $
Settlement	 3.44	0.177***	 0.50	0.021***	 3.93	0.197***
Immigration	 1.54	0.107***	 -0.36	0.022***	 1.19	0.086***
Relationships	 1.46	0.060***	 -0.63	0.025***	 0.83	0.035***
Jobs	 1.32	0.094***	 -0.41	0.026***	 0.91	0.067***
Cost of Living	 1.18	0.071***	 0.51	0.027***	 1.69	0.098***
Healthcare	 1.01	0.050***	 -0.35	0.016***	 0.66	0.033***
Finance	 0.92	0.061***	 0.35	0.020***	 1.27	0.081***
Education	 0.80	0.057***	 -0.46	0.031***	 0.34	0.026***
Housing	 0.52	0.058***	 0.26	0.023***	 0.78	0.081***
Transport	 0.42	0.025***	 0.53	0.029***	 0.95	0.054***
Diaspora	 0.07	0.002	 0.13	0.004	 0.20	0.006
Politics	 -0.05	0.003	 0.53	0.033***	 0.48	0.030***
Media/Food	 -1.87	0.051***	 0.38	0.011**	 -1.49	0.041***
Tech	 -4.04	0.112***	 -0.38	0.011***	 -4.42	0.123***
Gaming	 -6.71	0.158***	 -0.60	0.015***	 -7.31	0.173***

have been navigated, consistent with accounts of immigrant political incorporation as a gradual process that unfolds after settlement rather than upon arrival.

Pre-arrival to post-arrival. The cumulative pre-arrival to post-arrival comparison shows that migration produces a durable reorganization of online attention rather than a transient arrival-period spike. Settlement-geography discussion has the largest sustained increase in the table, followed by cost of living, finance, immigration procedures, transport, jobs, relationships, housing, healthcare, and politics, with education showing a modest sustained increase as well. The categories that remain depressed over the long term are gaming and technology, which together show the two largest negative changes in the table, alongside a smaller sustained decline in media-and-food discussion. What this profile makes visible is that the displacement of leisure and general-interest discourse by settlement-instrumental discourse is not a temporary feature of the arrival period but a long-lasting reorganization of how immigrants in our cohort spend their online attention. This durability is consistent with work on administrative burden, which characterizes interactions with public systems as cumulative rather than episodic and as generating learning, compliance, and psychological costs that persist over time (Moynihan, Herd, and Harvey 2015), and it aligns with prior findings that immigrants treat platforms like Reddit as ongoing settlement infrastructure rather than a one-time information source (Monteiro 2024).

A separate observation that emerges across all three transitions concerns diaspora-oriented discussion. Diaspora is the only topic in the table that does not reach statistical significance in any phase comparison. This is inter-

pretively meaningful because it suggests that, for the cohort represented in our data, migration is articulated less through identity-oriented or transnational discourse and more through practical, economic, and infrastructural concerns. This is consistent with the demographic profile of immigrants who are most visible on Canadian Reddit communities, who tend to be younger, more digitally engaged, and disproportionately drawn from economic-class and study-to-work pathways (Barthel et al. 2016), and for whom the most pressing dimensions of settlement are typically instrumental rather than identity-based.

Summary. Three observations consolidate this section. First, the topical reorganization around migration is structurally sharper than the affective reorganization documented in RQ1, with effect sizes an order of magnitude larger and a directional pattern that runs across most categories. Settlement geography, immigration procedures, jobs, finance, healthcare, housing, education, transport, and cost of living rise together within the same individuals, while gaming, technology, and media-and-food decline. Second, this reorganization is concentrated at the arrival boundary itself. The pre-arrival to arrival transition produces the largest and most coordinated shift, after which some topics recede while others continue to rise as immigrants move from acute settlement into embedded daily life. Third, the long-term topical profile is dominated by practical and infrastructural concerns rather than by identity-based discourse, with diaspora the only topic showing no significant phase shift, while cost of living, transport, housing, finance, settlement geography, and politics remain elevated long after arrival. The overall picture is one in which migration in our data is, above all, an informational and material reorientation, in which new-

comers' online attention is durably redirected toward the practical demands of building a life in a new country. Set against the affective and pronominal findings in RQ1, these results suggest that the language and the concerns of migration move in different registers. Affective change is modest and gradual, while topical change is sharp and structural, with the moment of arrival marking the clearest pivot point in immigrants' everyday online life.

Discussion

Our findings portray migration as a temporally extended and emotionally complex process, rather than a single discrete event. We discuss how our results contribute to understanding immigrant adaptation, and further situate our findings in broader migration literature.

Reddit communities and representational scope. Our findings should be interpreted in light of the demographic characteristics of Reddit users, who represent a specific and non-random subset of immigrants to Canada. Prior work consistently shows that Reddit users skew younger, male, educated, and more digitally literate than the general population, with disproportionately high participation among individuals working or aspiring to work in technical and professional sectors (Barthel et al. 2016). In the Canadian immigration context, this demographic profile aligns most closely with economic-class immigrants, international students, and temporary residents navigating education-to-work pathways, while underrepresenting older migrants, refugees, family-sponsored immigrants, and individuals with limited digital access or language proficiency. As a result, Reddit-based discourse is best understood as reflecting the experiences and concerns of immigrants who are both digitally engaged and institutionally exposed to platform-mediated information environments.

Migration as a process rather than an event. Our findings support a processual understanding of migration in which arrival marks the beginning, rather than the resolution, of settlement. Migration scholars have emphasized that settlement unfolds unevenly across economic, institutional, social, and affective domains over time (Ager and Strang 2008). Our longitudinal analyses provide empirical support for this view where affective and linguistic shifts intensify after arrival rather than at arrival, while topic-level dynamics show that institutional, financial, and material concerns persist well beyond entry. Schinkel (Schinkel 2018) argues that integration operates as a normative, state-centered logic that positions migrants as objects of evaluation and frames settlement as a linear movement toward an implicitly defined societal standard. While our findings likewise reject arrival as an endpoint, we do not interpret observed emotional or topical shifts as progress toward integration. Instead, we use longitudinal language data to document how post-arrival experiences remain structured by ongoing institutional, material, and social demands, without imposing a benchmark of successful incorporation.

Emotional adjustment and population-specific pathways. Our findings also need to be situated relative to prior

work on migration and mental health that focuses on populations with different migration pathways and structural exposures. Much of the evidence linking migration to heightened negative affect and psychological distress comes from studies of refugees and labor migrants facing acute displacement, legal precarity, or trauma (Lindert et al. 2009; Bhugra 2004). By contrast, the Reddit communities examined here are more likely to reflect younger, educated, and digitally engaged immigrants navigating education-to-work or economic migration pathways. This difference in population composition may help explain why we do not observe large increases in overt negative affect, even as post-arrival strain remains evident through diminished positive emotion, sustained social orientation, and shifts in discursive framing. Rather than contradicting prior findings, our results suggest that emotional adjustment pathways vary substantially across migrant subgroups, and that for relatively resourced migrants, post-arrival strain may manifest less through intensified negative emotion and more through prolonged negotiation of institutional, economic, and social demands.

Institutional burden and social media as settlement infrastructure. Across both affective and topical analyses, institution-facing concerns remain prominent well after arrival. Topics related to immigration processing, finance, housing, transportation, and geographic settlement persist across phases, indicating that post-arrival life is shaped by ongoing interaction with bureaucratic and infrastructural systems. This pattern aligns with work on administrative burden, which characterizes public services as generating cumulative learning, compliance, and psychological costs rather than one-time hurdles (Moynihan, Herd, and Harvey 2015). In migration contexts, these burdens are often intensified by legal complexity and uncertainty around eligibility and timelines. Prior research shows that immigrants increasingly rely on social media platforms as informal settlement infrastructure to manage these sustained demands, particularly in high-income destination countries such as Canada (Dekker et al. 2018; Leurs and Smets 2018; Veronis, Tabler, and Ahmed 2018). Consistent with this literature, our findings suggest that online participation is integral to post-arrival settlement, with platform-mediated interaction supporting collective navigation of institutional constraints over time.

Social orientation without assuming integration. Although social-process language and community-oriented topics increase after arrival, these shifts should not be equated with social integration or belonging. Prior work distinguishes between social participation and deeper forms of inclusion, noting that increased interaction may reflect necessity or dependence rather than incorporation (Anthias 2013; Putnam 2007). In our data, heightened social orientation co-occurs with reduced self-referential language and sustained engagement with institution-facing concerns, suggesting that post-arrival sociality is shaped by instrumental and informational demands. In digitally mediated settings, social interaction often supports advice-seeking and collective sensemaking without implying durable social ties (Leurs and Smets 2018).

Design, Policy, and Responsible Use

Our results also speak to broader questions about how social media data should and should not be used. On the one hand, longitudinal patterns in emotional and topical expression could help institutional actors, such as settlement agencies, universities, or municipal governments, anticipate when and where support is most needed. For example, spikes in negative affect and IRCC-related topics might indicate moments when backlogs or policy changes are causing widespread distress. On the other hand, any use of social media data for intervention must respect the autonomy and privacy of migrants. Our analysis is expressly aggregate and observational; we do not propose that platforms or institutions identify or monitor individual migrants based on their posts, nor that emotional markers be used for automated triage. Instead, we advocate for using insights at the cohort level to inform resource allocation, program design, and the timing of communication campaigns, while maintaining clear boundaries against individual-level profiling.

Limitations of This Study

Several limitations qualify our findings. **First**, our cohort consists of Reddit users who self-report as immigrants to Canada. This population is likely younger, more digitally engaged, and more comfortable with English than the broader immigrant population, which limits generalisability. **Second**, our arrival-year inference pipeline, while validated, is imperfect; some inferred years may be off by one year, especially near calendar boundaries, and non-standard temporal expressions may be misinterpreted. **Third**, LIWC-based markers, though interpretable and stable, offer a coarse view of emotional expression, and may not fully capture culture- and language-specific nuances in immigrant discourse. **Fourth**, our observational design cannot definitively separate the effects of migration from contemporaneous events (e.g., economic cycles, policy changes, or the COVID-19 pandemic) that also shape online expression. As such, our future work could address these limitations by combining social media with administrative or survey data, by comparing immigrant cohorts across multiple destination countries, and by integrating qualitative analysis with our quantitative findings. It would also be valuable to co-design tools with immigrants and settlement organisations, using our cohort-level insights as a starting point rather than an endpoint.

Concluding Remarks

This paper introduces a longitudinal analysis of immigrants' emotional and topical trajectories on Reddit as they move to and settle in Canada. By combining an LLM-based arrival-year inference pipeline with LIWC-based psycholinguistic markers and semantic clustering, we reconstruct pre-arrival, arrival-year, and post-arrival timelines for 9,834 self-reported immigrants with 4.8M posts.

Our findings show that migration reshapes online discourse less through dramatic emotional change than through a recomposition of self-positioning and a durable reallocation of topical attention. Affective categories move only

modestly, with only positive emotion declining substantively post-arrival, while the strongest within-user signal is pronominal, i.e., first-person singular falls sharply after arrival, collective *we* and external *they* rise across every phase, friend-related language declines, and family-related language modestly increases. Topical change is sharper at the arrival boundary, where settlement geography, immigration procedures, jobs, cost of living, healthcare, finance, education, housing, and transport rise together, while gaming, technology, and media/food decline. After arrival, acute procedural topics partially recede while cost of living, transport, housing, finance, settlement geography, and politics remain elevated; diaspora is the only topic showing no significant shift. Migration in our data is therefore best understood as an informational, relational, and material reorientation rather than a discrete emotional shock, with newcomers' attention durably redirected toward the practical demands of building everyday life in a new country.

References

1988. Canadian Multiculturalism Act. R.S.C., 1985, c. 24 (4th Supp.).
- Ager, A.; and Strang, A. 2008. Understanding integration: A conceptual framework. *J. Refugee Stud.*, 21: 166.
- Ahmed, S.; Jaidka, K.; Chen, V. H. H.; Cai, M.; Chen, A.; Emes, C. S.; Yu, V.; and Chib, A. 2024. Social media and anti-immigrant prejudice: a multi-method analysis of the role of social media use, threat perceptions, and cognitive ability. *Frontiers in psychology*, 15: 1280366.
- Anthias, F. 2013. Identity and belonging: Conceptualisations and political framings. *Nordic Journal of Migration Research*, 2(2): 102–110.
- Barthel, M.; Stocking, G.; Holcomb, J.; and Mitchell, A. 2016. Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center*, 25.
- Bhugra, D. 2004. Migration and mental health. *Acta psychiatrica scandinavica*, 109(4): 243–258.
- Carling, J.; and Collins, F. 2020. Introduction: Aspiration, desire and drivers of migration. In *Aspiration, desire and the drivers of migration*, 1–18. Routledge.
- Coppin, J. 2020. *The representations of migrants in articles shared on Reddit and their related comments*. Master's thesis.
- Dekker, R.; Engbersen, G.; Klaver, J.; and Vonk, H. 2018. Smart refugees: How Syrian asylum migrants use social media information in migration decision-making. *Social Media+ Society*, 4(1): 2056305118764439.
- Ding, C.; and He, X. 2004. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, 29.
- Garg, M.; Gaur, M.; Goswami, R.; and Sohn, S. 2023. Lost: a mental health dataset of low self-esteem in reddit posts.
- Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

- Immigration, R.; and Canada, C. 2025. Supplementary Immigration Levels Plan 2025 to 2027. <https://www.canada.ca/en/immigration-refugees-citizenship/news/notices/supplementary-immigration-levels-2025-2027.html>.
- Kamarudin, N. S.; Beigi, G.; Manikonda, L.; and Liu, H. 2020. Social media for mental health: Data, methods, and findings. In *Open Source Intelligence and Cyber Crime: Social Media Analytics*, 195–220.
- Leurs, K.; and Smets, K. 2018. Five questions for digital migration studies: Learning from digital connectivity and forced migration in (to) Europe. *Social media+ society*, 4(1): 2056305118764425.
- Lindert, J.; von Ehrenstein, O. S.; Priebe, S.; Mielck, A.; and Brähler, E. 2009. Depression and anxiety in labor migrants and refugees—a systematic review and meta-analysis. *Social science & medicine*, 69(2): 246–257.
- MacQueen, J. 1967. Multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297.
- Mittal, J.; Belorkar, A.; Jakhetiya, V.; Pokuri, V.; and Guntuku, S. C. 2023. Language on reddit reveals differential mental health markers for individuals posting in immigration communities. In *Proceedings of the 15th ACM Web Science Conference 2023*, 153–162.
- Monteiro, S. 2022. Social media and internet usage rates on employment outcomes among newcomers in Canada. Technical report.
- Monteiro, S. 2024. Searching for settlement information on Reddit. *International Migration*, 62(3): 100–119.
- Moynihan, D.; Herd, P.; and Harvey, H. 2015. Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*, 25(1): 43–69.
- Nguyen, T. T.; Mullaputi, P. S. P.; Yue, X.; Mane, H.; Santhosh, A.; Dennard, E.; Alibilli, A. S.; Makres, K.; McCoy, R. G.; and Nguyen, Q. C. 2025. A Decade of Discourse: Exploring Sentiments and Trends around Immigration on Social Media from 2014-2024. *Social Science & Medicine*, 118715.
- of Canada, G. 1971. Canada adopts an official multiculturalism policy. <https://www.clo-ocol.gc.ca/en/canada-adopts-official-multiculturalism-policy>.
- Phan, A. N. 2023. reddit_immigration: Analysis of subreddits on Canadian immigration programs. https://github.com/ngocmphan/reddit_immigration. GitHub repository.
- Putnam, R. D. 2007. E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture. *Scandinavian political studies*, 30(2): 137–174.
- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.
- Sabri, N.; Pham, A. C.; Kakkar, I.; and ElSherief, M. 2024. Inferring mental burnout discourse across Reddit communities. In *Proceedings of the Third Workshop on NLP for Positive Impact*, 224–231.
- Saha, K.; Kotakonda, B.; and De Choudhury, M. 2025. Mental Health Impact of the COVID-19 Pandemic on College Students: A Quasi-Experimental Study on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 1748–1770.
- Saha, K.; Torous, J.; Caine, E. D.; and De Choudhury, M. 2020. Psychosocial effects of the COVID-19 pandemic: large-scale quasi-experimental study on social media. *Journal of medical internet research*, 22(11): e22600.
- Saha, K.; Yousuf, A.; Boyd, R. L.; Pennebaker, J. W.; and De Choudhury, M. 2022. Social media discussions predict mental health consultations on college campuses. *Scientific reports*, 12(1): 123.
- Schinkel, W. 2018. Against ‘immigrant integration’: For an end to neocolonial knowledge production. *Comparative migration studies*, 6(1): 31.
- Stravato Emes, C. 2024. Is it about “them”? leveraging big data research to understand anti-immigrant discourse. *Big Data & Society*, 11(2): 20539517241249432.
- Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.
- Veronis, L.; Tabler, Z.; and Ahmed, R. 2018. Syrian refugee youth use social media: Building transcultural spaces and connections for resettlement in Ottawa, Canada. *Canadian Ethnic Studies*, 50(2): 79–99.
- Vitiugin, F.; Lee, S.; Paakki, H.; Chizhikova, A.; and Sawhney, N. 2024. Unraveling Code-Mixing Patterns in Migration Discourse: Automated Detection and Analysis of Online Conversations on Reddit. *arXiv preprint arXiv:2406.08633*.
- Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33: 5776–5788.
- Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Huang, J.; and Ananiadou, S. 2024. MentaLLaMA: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, 4489–4500.
- Yuan, Y.; Saha, K.; Keller, B.; Isometsä, E. T.; and Alavedood, T. 2023. Mental health coping stories on social media: a causal-inference study of Papageno effect. In *Proceedings of the ACM web conference 2023*, 2677–2685.
- Zikic, J.; Zupic, I.; and Černe, M. 2025. Digital nomads in conversation: Reddit-based analysis and the future of nomadic versus migrant career journeys. *International Migration Review*, 01979183241306343.

AAAI ICWSM Paper Checklist

Written by the AAAI ICWSM 2025 Organizing Committee

pc.chairs@icwsm.org

Abstract

This document offers a Paper Checklist to be appended at the end of all submissions to rounds of the AAAI ICWSM conference.

Paper Checklist to be included in your paper

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? No, we engage in this work to better understand the experiences of vulnerable migrants, and take care not to disclose the contents of the Reddit posts/comments.
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we discuss how our approach is justified in our Methods section, and discuss statistical methods in the Findings.
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we detail this in our Methods section.
- (e) Did you describe the limitations of your work? Yes, in the Discussions section, we have limitations of the study detailed.
- (f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss how we take particular care to address any potential negative impact of our work in our Discussion section
- (g) Did you discuss any potential misuse of your work? Yes, we discuss any potential misuse in our Discussion.
- (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? We anonymize our data and release

code and data responsibly to the best of our capacity. An Ethical Consideration section is added to the Appendix to further clarify this.

- (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? Yes, we provided as much detail as we could in both Methods, and Findings sections.
- (b) Have you provided justifications for all theoretical results? Yes, we provided as much detail as we could in both Methods, and Findings sections.
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, we discuss them in our Discussions section.
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, we account for some alternative explanations in our Discussion section.
- (e) Did you address potential biases or limitations in your theoretical framework? Yes, this is highlighted in our Limitations subsection.
- (f) Have you related your theoretical results to the existing literature in social science? Yes, we situate our results in broader literature on the topic in Discussions section.
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discuss this in our Discussions section.

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? N/A
- (b) Did you include complete proofs of all theoretical results? N/A

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)? Yes, we provide the code in our GitHub, but we do not disclose the Reddit posts/comments to ensure privacy.

- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we disclose information about models and parameters.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
 - (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)? Yes, we include it in the Appendix.
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, to the best of our capacity given the nature of the data.
 - (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? Yes, but cost is negligible and report our human annotation scores.
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, we provided all citations.
 - (b) Did you mention the license of the assets? All models used were open-source/open-weights and can be used for research freely.
 - (c) Did you include any new assets in the supplemental material or as a URL? No
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? N/A
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we discuss how we anonymized all posts and comments, and only work with aggregate data for analysis.
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? N/A
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? N/A
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? N/A
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A
 - (d) Did you discuss how data is stored, shared, and de-identified? N/A

References

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

Appendix

A: Ethical Consideration

Social Media and Data Ethics This study uses publicly accessible Reddit posts and comments published in open immigration-related forums. Data collection followed Reddit’s API terms of service. We did not access private or restricted content. In all analyses, we strip post IDs, URLs, and other direct identifiers; they are not included in any shared dataset or publication. Storage and processing took place in secured institutional environments. User-level segmentation is employed solely to reconstruct pre-arrival and post-arrival timelines and to compute aggregate statistics. Manual inspection is restricted to a small validation subset for pipeline evaluation, and in those cases identifiers are removed or hashed. Results are reported only in aggregate; we do not profile or describe any individual user. Because the data are public and contributed voluntarily, we did not obtain individual informed consent. This aligns with prevailing norms in computational social science for low-risk secondary analysis of public web data. However, we recognise that immigrants constitute a structurally vulnerable population, whose legal status, employment, and mental health may be sensitive. Our goal is not to diagnose individuals or infer personal outcomes, but to characterise population-level linguistic patterns around migration with the aim of informing more supportive systems and policies.

Mental Health Language Because our work touches on mental health-related language, we take care to avoid over-interpretation. LIWC features are used to detect patterns of emotional expression and potential distress in aggregate, not to assign clinical labels to individuals. Following the interpretive stance of Saha et al (Saha, Kotakonda, and De Choudhury 2025), we treat linguistic markers as indicators of mood or stress in population context rather than as diagnostic tools. In our analyses, we use these indicators to understand how migrants speak about uncertainty, stress, hope, adaptation, and belonging around the migration event. We do not attempt to predict self-harm risk, diagnose disorders, or trigger any automated intervention based on these markers. Instead, we focus on cohort-level trends.

Responsible Use of LLMs Our pipeline employs LLMs for two tasks, namely identifying self-reported immigrants and inferring arrival years from temporal expressions. All LLM inference is performed within our own research environment where we use a distilled version of Llama4 Scout, running on a locally hosted Ollama server that uses a single Nvidia RTX A6000 48GB. Text is not sent to external third-party services for processing, and no user-generated content is used to train models. We select relatively lightweight, open-weight models to reduce computational and environmental cost, and we cache intermediate outputs to avoid repeated passes over the same data. We deliberately confine LLM use to tasks where they provide clear added value (e.g., resolving temporal references) and where outputs can be validated or sanity-checked via manual annotation. LLM predictions are not treated as ground truth; instead, they are one component in a broader analytic workflow.

B: LLM Settings and Prompts

Model Specifications and Environment We use Llama-4-Scout-17B-16E-Instruct (GGUF, Q2_K_XL quantization), a 17B-parameter sparsely activated mixture-of-experts model, deployed via Ollama, with inference executed on a single NVIDIA RTX A6000 (48 GB VRAM) on our local lab server.

The prompts used for annotation are provided below.

Prompt 1: Single-Post Immigrant Classification

You are an AI annotator. Your task is to annotate Reddit posts and comments, and find if the user is an immigrant to Canada.

Return the result in JSON in the following format.

Output (JSON):

```
{
  "is_immigrant": {
    "value": "yes"|"no"|"maybe",
    "explanation": "string"
  },
  "confidence": "high"|"medium"|"low"
}
```

is_immigrant

• Value

- **“yes”**: Clear first-person immigrant narrative (e.g., “I moved to Canada in 2020”), and immigrated after 2010.
- **“no”**: News, general advice, third-party stories, or long-term residents (e.g., moved before 2010).
- **“maybe”**: Ambiguous but potentially immigrant-related content. Use “maybe” if:
 - * Use when there is indirect evidence or contextual clues that suggest a possible immigrant status, even if not directly stated.
 - * The post mentions arriving in Canada but lacks clarity on permanence, or suggests temporary stay (e.g., layovers, travel, vacation, etc.).
 - * If mentions first-person account of moving to Canada, but there is a reasonable chance the user has lived in Canada for a very long time (e.g., since childhood).
 - * If there is mention of an immediate family member of similar age (spouse, siblings) moving to Canada, there is a chance the user might be an immigrant too.

• **Explanation**: Provide a brief justification for the classification.

• Confidence Assessment

- **High**: Direct, clear statements (e.g., “I moved to Canada in 2020”).
- **Medium**: Can be reasonably inferred from context, but not explicitly mentioned cues.
- **Low**: Ambiguous mentions, mixed signals, or indirect references.

Reddit Post/Comment is given below:

```
{{reddit_post}}
```

Prompt 2: Author-Level Immigrant and Arrival-Time Inference

You are given up to 10 Reddit posts and comments from the same author. Each item includes:

- subreddit name
- creation date in UTC (YYYY-MM-DD)
- text content

Your task is to:

1. Determine if this author is likely an immigrant to Canada.
2. If yes, determine **when** they arrived in Canada:
 - If the text contains an explicit year, return that year (YYYY).
 - If the text contains only a relative time expression (e.g., “2 years ago”, “last year”, “in 2019”), return the exact phrase as it appears in the text.
 - If there is no clear year or relative time expression, return Unknown.
3. Give a confidence level (High, Medium, Low) for your determination.
4. Briefly explain your reasoning based on the provided posts/comments.

Guidelines:

- The author is an immigrant to Canada if they state or strongly imply they moved/immigrated/arrived in Canada from another country, OR if there is strong contextual evidence (e.g., landed PR, first tax year, OHIP application, settlement services, etc.).
- Do not try to calculate the arrival year from relative expressions — just copy them exactly.
- Confidence should be based on the strength and clarity of the evidence.
- Multiple posts may contain different information — choose the most direct and credible reference for arrival timing.

Respond in this exact JSON format:

```
{
  "is_immigrant_to_canada": true | false,
  "arrival_time": "YYYY" | "relative phrase" | "Unknown",
  "confidence": "High" | "Medium" | "Low",
  "reasoning": "1-2 sentences explaining your decision."
}
```

Here is the data for this author:

{{AUTHOR_POSTS_AND_COMMENTS}}

C: Temporal Topic Trajectories Around Arrival

To complement the phase-level comparisons, we examine how topic prevalence changes year by year around the inferred arrival date. Figure 5 plots the share of posts assigned to each topic across relative years, where year 0 denotes the arrival year. This view helps distinguish topics that spike

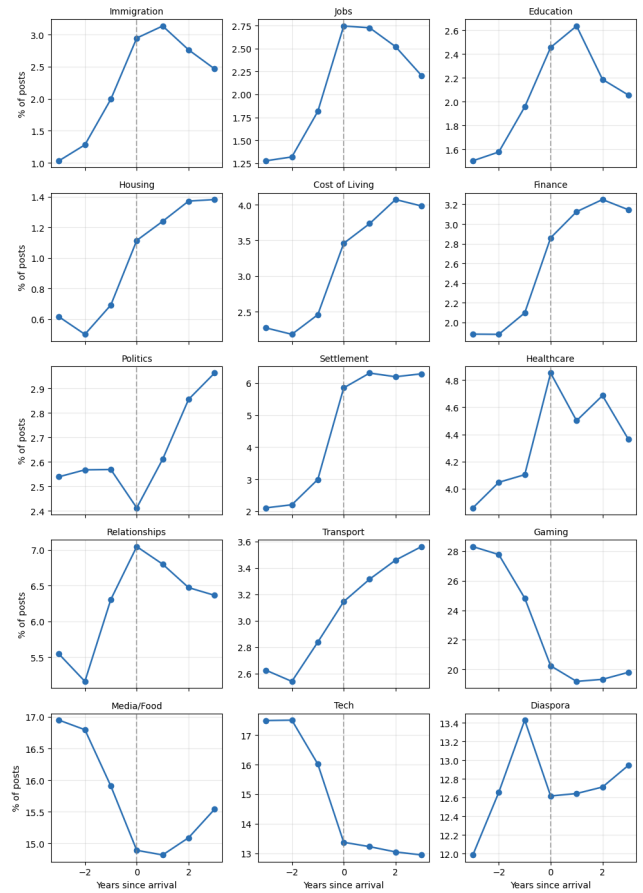


Figure 5: Topic trajectories by years since arrival. Each panel shows the percentage of posts assigned to a topic across years relative to the inferred arrival year, with year 0 marking arrival. The dashed vertical line indicates the arrival boundary. The trajectories show how immigrants’ topical attention shifts from pre-arrival discussion toward settlement, institutional, financial, and infrastructural concerns during and after arrival.

around arrival from those that remain elevated after settlement begins. The trajectories show a clear reallocation of attention at the migration boundary: settlement geography, immigration procedures, jobs, finance, housing, healthcare, cost of living, and transport rise around arrival, while gaming, technology, and media/food decline. Some topics, such as immigration and jobs, peak near the arrival year and then partially recede, whereas cost of living, housing, transport, finance, and politics remain elevated or continue to grow after arrival. These patterns suggest that migration produces not only a short-term informational shock around entry, but also a longer-term redirection of online attention toward the material and infrastructural demands of everyday life in Canada.

D: Anonymized Qualitative Examples

To support the interpretability of the topic labels used in our analysis, Table 4 provides three shortened representative examples for each final topic category. The examples are lightly shortened to protect the anonymity of the posters.

Table 4: Representative shortened examples

Topic	Representative shortened examples
Immigration & IRCC timelines	1. A user describes withdrawing and resubmitting an Express Entry application after long IRCC delays and uncertainty over eligibility streams. 2. A user asks whether leaving Canada while an inland PR application is processing is risky before receiving the PR card. 3. A student explains missing an IRCC biometrics request after university and family crises, resulting in refusal and uncertainty about restoring status.
Jobs, Skill-building & Career	1. A newcomer asks how to “dumb down” a consulting-heavy resume for entry-level or survival jobs. 2. A user describes being laid off from a first professional job, revising their resume, and treating job applications like full-time work. 3. A permanent resident considers whether to return to Canada after tech layoffs, high costs, and difficulty finding work.
Education & International Students	1. A prospective graduate student describes emailing professors without responses and shifting toward formal applications. 2. A student explains how course registration priority varies by program, specialization, honours status, and accumulated credits. 3. A user discusses English proficiency requirements, possible exemptions, and alternative accepted tests such as Duolingo.
Housing & Renting	1. A tenant worries that rising neighbourhood rents may motivate a landlord or agent to push them out. 2. A renter asks what a landlord must legally do to end a tenancy after relations have become tense. 3. A user describes severe pest problems in a Scarborough apartment and shares practical roach-control strategies.
Cost of Living & Consumption	1. A user asks about price ranges for clothing and whether cheaper but good-quality alternatives are available. 2. A commenter argues that reducing restaurant spending will not meaningfully solve housing affordability problems. 3. A user breaks down monthly income, spending, and investments while describing a low-expense lifestyle.
Personal Finance & Taxes	1. An executor and beneficiary asks how RRIF funds and high-interest savings interest will be handled for tax purposes. 2. A user is confused about TFSA contribution room after withdrawing from one TFSA and reinvesting in another. 3. A user checks CRA contribution limits and considers moving savings into a TFSA or Wealthsimple account.
Politics & Economy	1. A commenter discusses leveraged loans, CLOs, credit-rating failures, and parallels to the 2008 financial crisis. 2. A user argues that public debt and private debt should not be treated as the same in political debate. 3. A commenter debates progressive electoral strategy and the consequences of disengagement from party politics.
Settlement Cities	1. A user describes relocating to Cloverdale and finding the neighbourhood more pleasant than its reputation suggests. 2. A household plans to stay south of the river to avoid bridge commutes after experiencing a three-hour commute. 3. A commenter reflects on Toronto’s urban form and compares it unfavourably with European cities such as Amsterdam.
Geography &	
Healthcare & Medical Access	1. A user advises someone to confirm cancer stage and grade, research treatment options, and seek second opinions. 2. A commenter explains that a small mole or skin issue can often be treated quickly by a GP or walk-in clinic. 3. A user describes long ER and clinic waits, difficulty finding a family doctor, limited prescriptions, and specialist delays.

Continued on next page

Topic	Representative shortened examples
Relationships & Social Life	1. A user describes isolation after moving abroad, being unable to visit family in India, and wanting to connect with others. 2. A commenter encourages someone to focus on self-development, friendships, health, and career goals before major relationship decisions. 3. A user expresses fear about attending AIAnon while seeking support for a partner's relapse.
Transport & Driving	1. A user asks whether stop-and-go fuel consumption is systematically tested, especially for heavy vehicles. 2. A commenter analyzes traffic-light timing and crash video evidence involving drivers and a cyclist. 3. A user living in East Vancouver says biking and transit make car ownership unnecessary downtown.
Gaming & Sports	1. A player describes progressing quickly through a game build using specific items and league mechanics. 2. A user discusses receiving hate mail in PvP and the thrill of using a powerful temporary weapon. 3. A commenter gives puzzle-solving hints for a difficult Sudoku and recommends an online solving resource.
Media, Food & Cultural Consumption	1. A user recommends a Mexican butcher shop in Toronto for chorizo, groceries, and Mexican-style cuts of meat. 2. A commenter describes rediscovering Brussels sprouts after roasting them with oil, honey, and balsamic vinegar. 3. A user reflects on listening to <i>Illmatic</i> and asks for albums that immerse listeners in an artist's world.
Tech & Digital Tools	1. A user discusses batching small data grabs with Python and delivering them through a simple pipeline. 2. A user asks how to build an Android app that recognizes custom image patterns and extracts information from photos. 3. A commenter describes being trained to write PL/SQL scripts for extracting, validating, formatting, and loading data into cloud databases.
Diaspora & Non-English Communities	1. A commenter directs someone toward city-specific India forums and other trusted India-related travel resources. 2. A Spanish-language comment discusses harassment faced by university students during government internships. 3. A Telugu-language comment suggests creating a separate subreddit to sustain interest in Telugu-language discussions.

#	Keyword phrase
1	new immigrant to Canada
2	just arrived in Canada
3	moved to Canada
4	landed in Canada
5	landing in Canada
6	arrived in Canada
7	newcomer to Canada
8	newcomer in Canada
9	first week in Canada
10	first month in Canada
11	new in Canada
12	came to Canada
13	after getting PR in Canada
14	after landing as a permanent resident
15	new PR in Canada
16	international student in Canada
17	temporary worker in Canada
18	canada immigrant

Table 5: Keyword phrases used to retrieve candidate immigration-related Reddit posts and comments during data collection. The phrase “moved to Canada” appeared twice in the original list and is shown once here after de-duplication.

E: Keyword Phrases Used for Data Collection

Table 5 lists the keyword phrases used to retrieve candidate Reddit posts and comments for the initial immigration-related corpus. These phrases were selected to prioritize high-precision first-person migration narratives, especially posts in which users describe arrival, settlement, permanent residence, study, work, or newcomer status in Canada. Duplicate phrases were collapsed during search execution.