

SpanishTweetsCOVID-19: A Social Media Enriched Covid-19 Twitter Spanish Dataset

Antonela Tommasel

ISISTAN, CONICET-UNCPBA, Argentina
Johannes Kepler University Linz, Austria
Email: antonela.tommasel@isistan.unicen.edu.ar

Juan Manuel Rodriguez

Aalborg University, Denmark
ISISTAN, CONICET-UNCPBA, Argentina
Email: jmro@cs.aau.dk

Abstract—The COVID-19 pandemic reshaped how citizens, institutions, and governments communicated under crisis, exposing deep cultural, political, and emotional dynamics on social media. However, most analyses of these processes have focused on English-language data, leaving the Latin American public sphere underexplored. This paper introduces *SpanishTweetsCOVID-19*, a large-scale collection of over 278 million tweets related to the COVID-19 pandemic in Spanish language (primarily from Argentina) posted on *Twitter* between March 2020 and June 2021. By integrating official government communications with organic public discourse, the collection enables a fine-grained study of how information, emotions, and narratives evolved across Spanish-speaking communities, particularly in Argentina. Beyond serving as a data resource, the collection provides a foundation for investigating how communication strategies influence public engagement, the adoption of guidelines, the dynamics and spread of information, and the emergence of collective emotional patterns. It includes extensive metadata and psycholinguistic annotations that support cross-disciplinary analyses in computational social science, public health, and crisis communication. To foster transparent and reproducible research, the collection and accompanying tools are publicly available¹.

Index Terms—COVID-19, Social Media Data, Spanish language, Twitter, Infodemiology

I. INTRODUCTION

The COVID-19 pandemic led to an unprecedented increment of digital discourse, with social media platforms becoming central to public debate, governmental communication, and the spread of both accurate and misleading information. These platforms enable rapid message dissemination and two-way interactions, facilitating the exchange of critical information between citizens, governments, companies and institutions [1, 2]. Governments increasingly use them to engage citizens and involve them in decision-making processes [3, 4]. Effective communication requires monitoring citizens' diverse information needs and aligning messages with their values and experiences, particularly during crises [5]. Message framing also plays a crucial role in shaping behaviors, attitudes, and perceptions. While social media can help reduce uncertainty and anxiety by enabling the sharing of experiences, it also contributes to *infodemics*, an overwhelming flood of information, including misinformation and disinformation, that makes it harder to identify reliable sources [6]. The virality of

misleading content complicates crisis management, demanding strategies that not only promote socially beneficial behaviors but also mitigate the diffusion of narratives that can undermine public health efforts.

COVID-19 presented unique challenges for Latin America, a region marked by social and economic inequality, institutional fragility, and diverse health management capacities [7]. While extensive research has examined communication and behavior in the United States and Europe, Latin America has received comparatively less attention [8]. Nonetheless, cultural, political, and linguistic factors are central to understanding how citizens respond to official communication and how misinformation spreads. For example, in collectivist cultures, appeals to community well-being and solidarity tend to be more effective, while in individualist cultures, appeals to personal responsibility may have a greater impact [9, 10]. Also, levels of institutional trust vary widely across cultures, influencing the effectiveness of public communication [11, 10].

Despite the rapid growth of social media use across Latin America, few studies have systematically analyzed its role in crisis communication [11, 12]. In particular, large-scale, open collections capturing Spanish-language discourse from Latin American contexts remain scarce. This gap highlights the need for data reflecting diverse socioeconomic and cultural realities, where effective and inclusive communication can significantly affect crisis outcomes [13].

To address this gap, we introduce *SpanishTweetsCOVID-19*, a large-scale data collection of over 278 million Spanish-language tweets related to the COVID-19 pandemic, collected between March 2020 and June 2021, with a focus on Latin American discourse, particularly from Argentina. Argentina implemented early preventive measures even before detecting its first COVID-19 case and later intensified restrictions, including strict lockdowns, flight cancellations, border closures, and mandatory mask use [14]. Despite these efforts, by April 2020, Argentina had one of the highest infection rates, and by early 2021, ranked among the five most affected countries globally [11]².

¹The collection is available at: <https://data.mendeley.com/datasets/nv8k69y59d/5>.

²A detailed analysis of the political and social climate of Argentina during the pandemic, including how these dynamics were reflected in the communications and public discourse captured by this data collection, can be found in [15].

This data collection provides a unique opportunity to analyze how public communication and citizen discourse evolved during this period. Unlike prior collections focused mainly on English-language content, *SpanishTweetsCOVID-19* integrates organic public discussions with official government communications, enabling a comprehensive examination of crisis communication dynamics and their societal impact. It includes extensive metadata (such as, mentions, retweets, media, URLs, hashtags, replies, emotions, psycholinguistic categories, and content-based user relations) allowing for fine-grained analyses of interactions, engagement, and discourse evolution. By focusing on Spanish-language communication, the collection broadens the scope of research on social media’s role in crisis management, complementing prior studies predominantly based on English-language data [16].

In addition to the data collection, we provide a set of psycholinguistic resources specifically adapted for Spanish, designed to enhance the analysis of emotions, cognitive processes, and behavioral patterns in social media discourse. These resources include translated and adapted versions of established English-language lexicons. These enhance the collection’s analytical potential, enabling fine-grained studies of emotional evolution, mental health indicators, and communication framing in native Spanish text, areas often overlooked due to the scarcity of validated tools in this language. We also make available *Faking It*³, a companion tool for data rehydration and retrieval. The tool facilitates reproducible access to tweet content and metadata, complementing the collection’s analytical capabilities. Together, these resources aim to foster interdisciplinary research on crisis communication, misinformation, and emotional dynamics in Spanish-speaking contexts, expanding the methodological toolkit available to researchers working beyond English-language social media.

The remainder of this paper details the construction, structure, and potential applications of the collection, emphasizing its value for studying the societal impact of COVID-19 and, more broadly, for advancing crisis communication and digital behavior research in Latin America.

II. RELATED WORKS

Since the beginning of the COVID-19 pandemic, several data collections have been released, differing in scope, language coverage, and metadata availability. Although some collections are multilingual, most are dominated by English-language content, with Spanish tweets typically underrepresented and often limited to Spain. For example, Aguilar-Gallegos et al. [17] compiled a multilingual dataset of 8 million tweets collected over 23 days between January and February 2020 by searching for “Coronavirus” and its hashtag variant, capturing early discussions about the virus. Similarly, *TweetsCOVID19* [18] includes over 8 million tweets from October 2019 to April 2020, providing metadata such as entities, hashtags, user mentions, sentiment, and URLs. Lopez and

Gallemore [19] collected 2.2 billion tweets starting January 22, 2020, with Spanish tweets comprising 13%, mainly from Spain. English and Spanish tweets were enriched with sentiment analysis and named entity recognition. Similarly, Chen et al. [20] gathered 123 million tweets from January 21 to May 11, 2020, using trending hashtags like #Coronapocalypse and #PanicBuying to capture early pandemic discourse. While English dominates the collection (over 60% of tweets), Spanish tweets represent 11%, primarily from Spain. This collection includes only tweet IDs sorted by date, without metadata. Similarly, *GeoCoV19* [21] focused on geotagged tweets collected over a period of 90 days since February 1, 2020, including 5.4 million tweets with place information, and 378K geotagged tweets with GPS coordinates.

Resources centered on Spanish tend to focus on narrow topics or restricted time spans. Martínez et al. [22] shared two small stance-detection collections containing between 2.8k and 11k manually and machine-learning annotated tweets related to COVID-19 vaccination, primarily from Spain. Catalan-Matamoros et al. [23] also examined vaccine-related discussions across English, French, Portuguese, and Spanish-speaking communities. The collection comprised 3.7 million tweets, with Spanish tweets (primarily from Spain) accounting for 30% of them. Although these collections enable targeted analyses of topic modelling, and sentiment and risk perception, their limited scope prevents a broader understanding of crisis communication dynamics and cultural variability across Spanish-speaking contexts. In contrast, Alqurashi et al. [24] and Haouari et al. [25] focused on Arabic-language tweets. Alqurashi et al. [24] collected over 3 million Arabic tweets from January 1, 2020, tracking COVID-19-related keywords, while Haouari et al. [25] compiled 748k popular tweets from January to March 2020, along with their propagation networks. These collections illustrate the analytical value of region-specific datasets that capture local linguistic, cultural, and institutional realities, an approach still largely missing for the Latin American Spanish-speaking world.

Within Latin America, several studies analyzed social media discourse during the pandemic, but most did not make their data publicly available, hindering reproducibility and broader analyses of engagement, trust, and information diffusion. Yum [26] examined influential *Twitter* users across Brazil, Peru, Colombia, Chile, and Argentina, showing that governmental accounts were dominant in Peru, Chile, Argentina, and Ecuador, while news media led in Brazil and Colombia. This highlights the importance of including official government accounts in data collections. Amado [27] and Baumann and Denardi [28] conducted qualitative analyses on state-sponsored misinformation and Sinophobia, respectively, both relying on small, manually selected tweet samples. Arias et al. [29] analyzed public sentiment in Panama based on 26 million COVID-19 tweets from 2019 and 28 million from 2020 shared by manually selected authors, excluding government officials. Cabezas et al. [30] studied emotions in Spanish-speaking countries (Chile, Mexico, Peru, and Spain) over 13 months (March 2020–March 2021), analyzing 3 million tweets. Latin

³<https://github.com/knife982000/FakingIt>. Currently, the tool supports URL-based downloads and direct access to tweet content.

American tweets accounted for 50% of the collection, with Mexico contributing 30%. Demonte et al. [31] quantitatively examined 700 manually selected vaccine-related tweets from government, media, and other sources. Similarly, Córdoba-Cabús et al. [32] analyzed vaccine discourse in Argentina, Colombia, Chile, Mexico, and Peru, in 24k tweets from the five most influential media outlets per country collected between January 2020 and June 2023. Ceron et al. [33] explored misinformation in Latin America, analyzing 100k tweets shared between January and July 2020 by six major fact-checking agencies. Finally, Heredia [34] explored how whiteness and exclusionary practices were reproduced in online discourse, focusing on reactions to prisoner releases in Argentina due to COVID-19. Their study analyzed 400 manually selected tweets shared between January and July 2020.

In this context, our *SpanishTweetsCOVID-19* collection contributes a unique and comprehensive resource that bridges these gaps by offering both scale and analytical depth, capturing not only citizen discussions but also official governmental communication, thereby enabling the study of top-down and bottom-up information flows. Unlike prior collections that primarily focus on multilingual content with an English-language dominance, *SpanishTweetsCOVID-19* emphasizes Spanish tweets, ensuring a more representative analysis of discourse within Spanish-speaking communities, in particular, Argentina. Moreover, it integrates psycholinguistic and emotional annotations specifically adapted for Spanish, which support nuanced examinations of emotion, cognition, and communication strategies throughout the pandemic. By combining linguistic diversity, institutional coverage, and temporal richness, this collection moves beyond descriptive data sharing to serve as a foundation for interdisciplinary research on crisis communication, misinformation, and collective behavior in Spanish-speaking societies.

III. MATERIALS AND METHODS

A. Data collection process

The raw data was collected via the *Twitter* API between March 15, 2020 and June 4, 2021 using our custom tool, *Faking it!*, which integrates *Twitter4J* for seamless API access. Data was gathered through *Twitter*'s Streaming service, enabling real-time tweet collection. As detailed in Table I, filtering criteria required tweets to be in Spanish, contain at least one selected keyword, originate from within Argentina's bounding box or interact (e.g., mention, reply, quote, retweet) with an official Argentinian government or media account. This process yielded 278,307,942 tweets, including original posts, retweets, quotes and replies.

Social media platforms support two main types of user relationships: structural and conversational. Structural relationships, such as *Twitter*'s follower-followee connections or *Facebook* friendships, typically involve known acquaintances (e.g., colleagues, family, classmates) or individuals with shared interests. These connections tend to be reciprocal and long-term [35]. In contrast, conversational interactions focus on the shared and consumed content, making them more dynamic and

independent of structural ties. As a result, they provide a more timely reflection of users' interests. Conversational networks have been widely used to study filter bubbles, echo chambers, community structures, and the spread of misinformation and political content [36, 37], among others.

Given the dynamic nature of content-based relationships [38] and the direct availability of follower-followee data via the *Twitter* API, we focused on the conversational network surrounding COVID-19 discussions. We constructed a directed graph where nodes represent users and edges represent interactions (reply, mention, quote, or retweet). Following the flow of information, an edge from user *A* to user *B* indicates that *A* replied to, mentioned, or quoted *B*, whereas an edge from *B* to *A* indicates that *A* retweeted *B* [39, 36, 37].

Since raw data cannot be publicly shared, we created filtered tables, detailed in the Mendeley repository. In addition to tweet IDs, the collection includes information on mentions, retweets, media, URLs, hashtags, replies, users, content-based user relations and psycholinguistic resources summarized by day. All tables in the collection include the attribute *tweet_id*, enabling cross-table linkage and integration of information. Some tables focus on original content, excluding retweets to avoid duplication of shared messages, whereas others retain them to preserve the full interaction network and diffusion patterns. In particular, tables containing temporal and geographical information, as well as those defining user relationships, include retweets to support the reconstruction of key tweets and their propagation dynamics.

Code for processing and analyzing both raw data and shared tables is available in the *Faking it!* repository and the companion repository⁴. Given access to the *Twitter/X* API, the dataset can be rehydrated using tools like *Twitter4J V2*⁵ (Java), *Tweepy*⁶ (Python), or *Faking it!* (Java/Kotlin).

B. Psycholinguistic and Emotional Resources

To complement the data collection, we share a set of psycholinguistic resources that enhance the analytical potential of the *SpanishTweetsCOVID-19* collection. These include a Spanish adaptation of the *Empath* lexicon, which captures topical and emotional categories in text, and its adaptation to crisis and mental health lexicons. Together, these resources enable large-scale analyses of cognitive, emotional, and thematic patterns in Spanish-language social media.

1) *The Spanish Empath lexicon*: The *Empath* lexicon [40, 41] maps text into nearly 200 human-validated emotional and topical categories derived from *ConceptNet* and Parrott [42]'s hierarchy of emotions. These categories have been shown to correlate strongly with similar constructs in *LIWC* and *EmoLex*, making *Empath* a widely used tool for computational psycholinguistic analysis.

The adaptation of *Empath* for Spanish-language analysis was originally developed in Tommasel et al. [15], and here we provide a more detailed account of that process. Categories

⁴<https://github.com/tommantonela/SpanishTweetsCovid19>

⁵<https://github.com/takke/twitter4j-v2>

⁶<https://www.tweepy.org/>

TABLE I
Twitter STREAM FILTERING CRITERIA

language	es
keywords (and hashtags)	quedateencasa (stay at home), covid, covid-19, covid19, casa rosada (president's official workplace), cuarentena (quarantine, lockdown), barbijo (face mask), máscara (mask), salud (health), coronavirus, corona, solidaridad, argentina, muerto (dead, male), muerta (dead, female), muerte (death), infectado (infected, male), infectada (infected, female), test, testeo (testing), testeomasivo (massive testing), tapaboca (face mask), medico (md), enfermera (nurse), viruschino (chinese virus), virus, virus chino (chinese virus), pandemia (pandemic), mayorescuidados (more care), tapatelaboca (cover your mouth), cuidarteescuidarnos (taking care of yourself is taking care of us), desarrollo social (social development ministry), cancilleria argentina (Chancellery, referring to the ministry of foreign affairs), argentinaunida (united Argentina), solidaridad (solidarity), argentina, vacuna (vaccine), vacunatoriovip (vip vaccination), vacunagate (vaccine gate), sputnik, sinopharm, sinovac, pfizer, astrazeneca, vacunados vip (vip vaccination), abranto (let's open everything), cierrrentodo (let's close everything), fase1 (phase 1), cuarentenaestricta (strict lockdown), restricciones (restrictions), unidosporargentina (united by Argentina)
location	Argentina's bounding box (-73.4154357571, -55.1850761, -53.628348965, -21.8323104794)
users	• <i>government</i>: msalacion, casarosada, gcba, DiputadosAR, HCDiputadosBA, MDSNacion, Migraciones_AR, MindefArg, MindeTransporte, MinSeg, msnarg, MinGenerosAR, MinInteriorAR, MinTrabajoAR, cancilleriaarg • <i>politicians in office</i>: alferdez, kicillofok, felipe_sola, cfkargentina • <i>politicians</i>: patobullrich, mauriciomacri • <i>media</i>: LANACION, infobae, clarincom, c5n, elcancillercom, filonewsOK, Chequeado (fact-checking agency)

and associated terms were automatically translated using the IBM Watson Language Translator, and the resulting translations were independently reviewed to ensure consistency and accuracy. To address the gendered nature of Spanish, the lexicon was expanded to include gender-specific and, when appropriate, gender-neutral word forms.

This process follows best practices for linguistic and psychometric translation by Perczek et al. [43], who analyzed the reliability of Spanish translation from English for psychometric scales and lexicons. While an alternative would have been to translate tweets into English, the size of the dataset made this option impractical even with automated tools. By offering a validated Spanish version of *Empath*, we enable the large-scale psycholinguistic analysis of Spanish-language social media, facilitating studies on emotion, discourse framing, and collective behavior during crises.

2) *Empath for crisis and mental health analysis*: The language people use on social media to share opinions, experiences, and reactions offers valuable insights into their psychological states and collective emotions, particularly during crises. Lexicon-based approaches have proven effective for detecting emotional and cognitive signals in large-scale social data [15]. In this context, our Spanish-adapted *Empath* provides a framework for examining crisis communication and mental health discourse across time and societal actors [15].

Crises typically unfold across four interconnected stages [44, 45, 46]: preparedness, response, recovery, and mitigation. Each stage is characterized by distinct actions, sentiments, emotions, concerns, and perceptions from various actors (e.g., individuals, society, government), all reflecting in distinct linguistic markers. Language thus reflects both individual and collective reactions as crisis progress. Prior research has shown that lexicons can capture the nuances of crisis stages [47] and health indicators such as depression, stress, or anxiety in social media text [48, 49, 50].

Building on prior research [47, 48, 49, 50], Tommasel et al. [15] mapped existing English-language lexicons related to crises and mental health indicators to *Empath* categories based on lexical overlap. The most representative categories for each crisis stage or mental health indicator were identified

through this mapping. Following [15], tweet were considered to match with an *Empath* category if they contained at least one of the category associated words, disregarding repeated occurrences to account for tweet-length limitations. Finally, the daily prevalence of each category was calculated as the percentage of tweets per day containing terms from that category.

We acknowledge that this lexical-matching approach (similar to methods employed by tools such as *LIWC* and other lexicon-based frameworks) cannot fully capture the nuanced semantics or contextual subtleties of individual tweets. Nevertheless, our goal is not to achieve high precision at the single-tweet level, but to identify aggregated patterns and temporal fluctuations in emotional and thematic signals. At this scale, the volume of data helps smooth individual-level variability, allowing robust population-level inferences. This assumption aligns with prior research demonstrating that lexicon-based methods, despite their simplicity, can reliably capture collective behavioral and emotional trends when applied to large social media corpora [51, 15].

3) *Emotion analysis with SentiSense*: Sentiment analysis is a valuable tool for detecting social media content that enhances situational awareness by providing insights into individual dynamics [51], including how people cope with the causes, effects, and emotional burden of a crisis. While global sentiment scores provide useful indicators of public mood, they often overlook the complexity of emotional responses to crises. Individuals' reactions encompass a wide range of feelings linked to social, economic, and psychological factors [51]. Therefore, assessing multiple emotional dimensions offers a more comprehensive understanding of how individuals experience and respond to collective disruptions [52, 15].

To capture this diversity, we also incorporated the *SentiSense Affective Lexicon* [53, 54], which classifies *WordNet*⁷ synsets into 14 emotional categories based on classical models of affect [55, 56]. Since *SentiSense* is integrated with the Spanish version of *WordNet*, it can be directly applied to analyze the emotional content of tweets written in Spanish.

⁷<https://wordnet.princeton.edu/>

SpanishTweetsCOVID-19 includes the outputs of both lexicon-based analyses: one file reporting the daily prevalence of *Empath* categories (capturing psycholinguistic and thematic dimensions), and another reporting the daily distribution of *SentiSense* emotions. Together, these complementary resources enrich the data collection with interpretable features that link language use to emotional and cognitive dimensions of public discourse. They allow researchers to explore not only the polarity of public discourse but also its emotional richness and temporal evolution, providing a multidimensional view of how Spanish-speaking communities articulated fear, hope, anger, or solidarity throughout the COVID-19 crisis.

C. Data collection and analysis tool

Faking it! is a data collection and analysis tool designed to study the evolution and propagation of online content. Originally developed to trace how news and social media narratives emerge, spread, and transform over time, the tool enables researchers to retrieve web content from specified URLs and to monitor how posts referencing these sources circulate across social media platforms, including their retweet and reply chains. In addition to these diffusion trajectories, *Faking it!* collects accompanying URLs, images, and user metadata (such as profile information, posts, and social networks) thereby supporting detailed analyses of user engagement and information flow.

A key feature of *Faking it!* is that the web content extraction component works independently of any API, allowing the extraction of content and news-related information directly from URLs. Its flexible retrieval options support the collection of specific subsets, such as news content, related media, social media interactions, user metadata or full conversation threads, depending on the research goal.

By combining enriched metadata with temporal information, *Faking it!* facilitates the application of diverse computational and network-based techniques to detect, trace, and model the evolution of online narratives. This versatility makes it a practical resource for studying information diffusion, engagement dynamics, and the interplay between news and social media ecosystems.

D. Data value

The SpanishTweetsCOVID-19 collection provides a large-scale sample of social media discourse shared during the COVID-19 pandemic, offering a comprehensive view of *Twitter/X* across Spanish-speaking countries, with a particular focus on Argentina. Designed to capture both institutional and public communication, it includes posts from government offices, media outlets, and citizens, enabling the study of how official messaging, media narratives, and user-driven trends interacted and evolved throughout the crisis.

With over 278 million Spanish-language tweets, SpanishTweetsCOVID-19 is among the largest publicly available datasets of its kind. Beyond tweet content, it incorporates social and temporal metadata (such as retweets, mentions, quotes, and replies) allowing for the analysis

of interaction networks, echo chambers, and content propagation⁸. The collection also supports psycholinguistic and emotional analysis through Spanish-adapted lexicons, providing an additional layer for examining collective emotions, communication framing, and behavioral responses.

Researchers from multiple disciplines can leverage this resource to explore public engagement, misinformation dynamics, risk perception, and adherence to health guidelines. It is also valuable for developing and evaluating computational methods for crisis informatics, sentiment and emotion detection, and discourse evolution. The insights derived from such analyses can inform evidence-based communication strategies and support policymakers in understanding and responding to public concerns during health emergencies.

IV. DATA ANALYSIS

A. Data collection statistics

Table II and Figure 1 provide an overview of the dataset. Tweet statistics related to media, URLs, mentions, contributors and hashtags exclude retweets to avoid redundancy. Figure 1 shows the monthly distribution of tweets by type (original, retweets, replies and quotes). Although data collection began on March 15, 2020, the date of the first official communications from the Argentinian government, tweets from March represent less than 1% of the total, and are omitted in the Figure (the same applies to the tweets collected during June 2021). As the collection includes replies/quoted/retweeted posts, some earlier tweets are also included in the collection.

Overall, retweets dominate the collection (70%), followed by original tweets (20%) and replies (10%), illustrating the highly participatory and amplifying nature of *Twitter* discourse during the pandemic. Tweet volume peaked in April and May 2020, reflecting the heightened public attention surrounding the first lockdown measures, then gradually declined before rising again in October as new restrictions were announced. Weekly statistics reveal additional activity spikes in late April (week 17), late May (week 22), and late August (week 36), each coinciding with major government announcements related to confinement and economic policy.

The collection contains a diverse range of hashtags, highlighting both coordinated messaging and spontaneous user activity. The most frequent hashtags were #QuedateEnCasa (*stay at home*, 981,261 occurrences), #COVID19 (897,924) and #coronavirus (756,184). A total of 1,263,093 hashtags appeared only once, and 75% were used three times or fewer, revealing a long-tail distribution typical of organic social media activity. Variants in capitalization and spelling (e.g., #COVID19 vs. #Covid19) further reflect the heterogeneity of user-generated content.

User activity followed a similarly skewed pattern. Among the 14,901,562 users in the dataset, the median posted just three tweets, and 75% shared eight or fewer. At the same

⁸The collection has already been used for analyzing [15] and forecasting [57] mental health and crisis trends, and for analyzing communication intents and engagement drivers [58].

TABLE II
STATISTICS OF THE COLLECTED DATASET

Tweets	Start date: 15 March 2020		End date: 4 June 2021	
	Original: 57,637,586	Retweets: 192,398,361	Replies: 28,271,995	Quotes: 18,824,146
	With media: 18,774,720			
	With URL: 24,538,126			
	With mentions: 126,610,277			
	With hashtags: 15,487,896		Avg. number of hashtags per tweet: 2.09 ±2.05 Avg. frequency of hashtags: 16 ±1227 Min: 1 - Max 981,261	

Users	Total: 14,901,561		Verified: 50,330	
	Avg. tweets per user: 18.67 ±126.07		Min: 1 Max: 40,676	

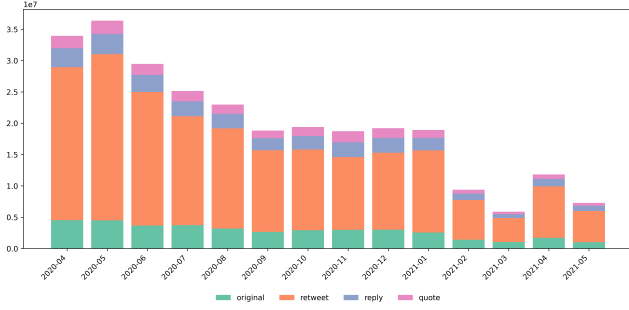


Fig. 1. Distribution of tweet types per month and type

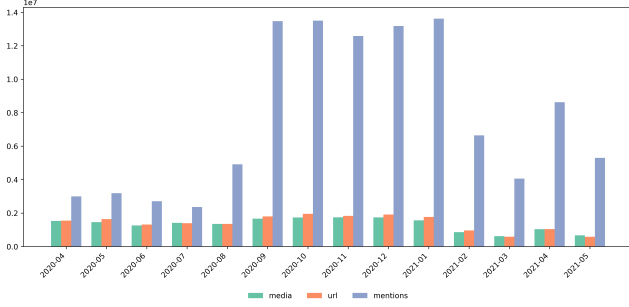


Fig. 2. Distribution of tweets with media content per month and type

time, a small subset of highly active users contributed disproportionately: over 171,756 users posted more than 250 tweets, 28,361 shared over 1,000, and 152 exceeded 10,000 tweets. The most prolific account published 82,606 tweets, with the most active users primarily belonging to digital media outlets, reinforcing their central role in shaping the online information landscape.

Finally, Figure 2 depicts the distribution of tweets by content type. Mentions were the most common element, appearing nearly five times as often as other features, followed by URLs and media content. The frequency of URLs, and media remained relatively stable across months, averaging about one per tweet. This consistency suggests persistent patterns of interaction and content sharing throughout the pandemic, even as overall activity levels fluctuated.

B. Psycholinguistic categories

To illustrate the analytical potential of the dataset, Figure 3 presents the temporal evolution of *Empath* categories associated with anxiety (*sadness*, *nervousness*, *fear*, *suffering*, *horror*, *disappointment*, *health*, *confusion*, *shame*, and *anger*) as defined in Tommasel et al. [15]. The analysis focuses on

the first part of the initial wave of the pandemic (March–June 2020)⁹. The period from July to October 2020 was excluded due to a sharp increase in marker prevalence around the 100-day lockdown milestone (late June), which hindered the analysis of earlier trends.

Darker regions in Figure 3 indicate higher category prevalence. A first wave of activity emerged following Argentina’s first confirmed COVID-19 case and in the days preceding the lockdown announcement (March 7–14). Around March 8, clear increases were observed in *confusion*, *horror*, *disappointment*, and *nervousness*, while *suffering* and *health* showed more gradual changes. A pronounced spike in *anger* during early March coincided with controversial statements by Argentina’s Health Minister minimizing the threat of the virus and political debates over expanded presidential powers. These dynamics mirror patterns observed by Aiello et al. [59], who reported similar surges in anger, fear, and anxiety in the U.S. following its first confirmed COVID-19 case, with subsequent stabilization after lockdowns were enacted. Additional peaks in our data correspond to key national events, including debt negotiations affecting the Argentine peso and public protests against government measures.

Figure 6 complements this analysis by showing the distribution of anxiety-related *Empath* categories over time. The relative prominence of these categories shifted notably across months. In March, *health* exhibited the highest prevalence and remained dominant through June, followed by a rise in *suffering*. From July onward, *confusion* became the most frequent category despite having shown the lowest levels earlier in the year. Similarly, *sadness*, which was initially concentrated around the upper quartile (Q3), shifted closer to the median (Q2) after June. Overall, anxiety-related language was less prevalent in the early months but increased substantially from July onward.

Figure 4 illustrates the evolution of *Empath* categories corresponding to the first three classical crisis stages (*preparedness*, *response*, and *recovery*) as defined in Tommasel et al. [15] and based on Neal [60] and Drabek [44]. The mitigation stage was excluded, as it pertains to post-crisis measures beyond the first wave period. Similar to the anxiety markers, changes in category magnitudes after July 2020 made earlier patterns less

⁹Additional visualizations covering the full collection, alternative time-frames, and other mental health dimensions (e.g., depression and stress) are available in the companion repository.

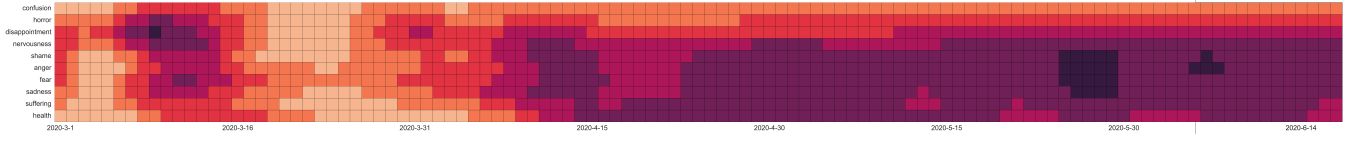


Fig. 3. Prevalence of the *Empath* categories related to mental health manifestations of Anxiety

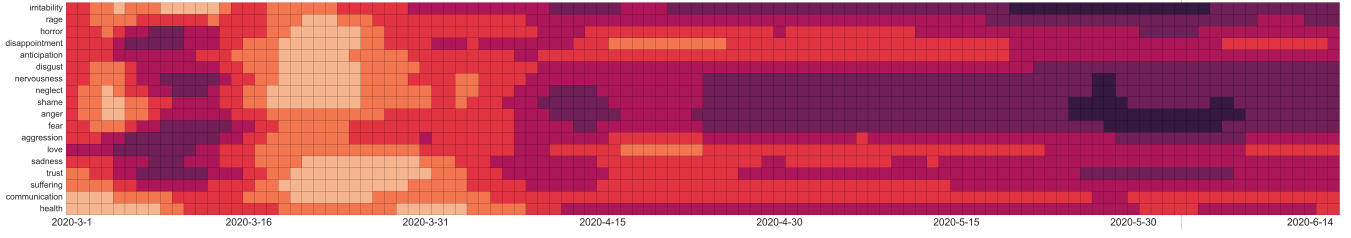


Fig. 4. Prevalence of the *Empath* categories related to the crisis stages

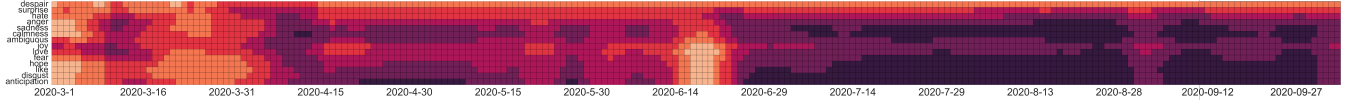


Fig. 5. Prevalence of the *SentiSense* emotions

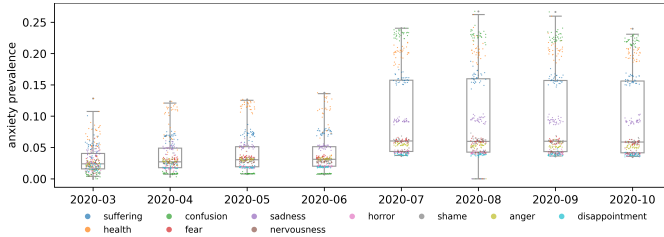


Fig. 6. Distribution of the *Empath* categories related to mental health manifestations of Anxiety per month

discernible. An increase in all three stage-related categories in late June paralleled a resurgence in case numbers, fatalities, and positivity rates, alongside worsening economic conditions caused by extended shutdowns. These findings suggest that category prevalence can act as an indicator for estimating the onset, duration, and overlap of crisis phases, insights that could inform communication strategies tailored to evolving public needs during different stages of a crisis.

An important methodological direction involves the analytical detection of *change points*, moments when category prevalence shifts significantly from prior levels, such as the darkened regions around early to mid-March. Building on approaches from Aiello et al. [59] and Tommasel et al. [15], automated change-point detection offers a systematic complement to visual interpretation of temporal variations. Once identified, these points can serve as anchors for forecasting models that predict future high-prevalence periods over short- or medium-term horizons [57]. Such predictive capabilities would enable decision-makers to anticipate emotional trends, evaluate communication effects, and plan timely interventions to mitigate crisis impacts.

Finally, the prevalence of linguistic categories (or other phenomena like communication intents) can also be applied to analyze communication patterns on social media, particularly

those of government accounts [58]. Examining these linguistic trends in relation to user engagement metrics, such as popularity (likes), virality (retweets), and commitment (replies), can reveal how institutional communication evolves with the crisis and adapts to shifting citizen needs. This type of analysis can inform the development of data-driven communication strategies that foster trust, engagement, and effective crisis response.

C. Emotions

Figure 5 illustrates the evolution of emotional expressions from the confirmation of Argentina's first COVID-19 case to the announcement of the initial lockdown in early to mid-March 2020. During this period, both negative emotions (*despair, hate, anger, sadness, disgust*) and positive emotions (*surprise, calmness, joy, love, hope, anticipation*) increased markedly. The surge in positive emotions likely reflects initial trust in government measures and a collective sense of reassurance, while the rise in *anticipation* signals growing uncertainty about the upcoming restrictions. By early June (when lockdown was extended for the sixth time but partially relaxed in major urban areas to allow outdoor activities and limited commerce) overall emotional expression declined, suggesting a brief phase of cautious optimism. However, as cases surged again in late June and another extension was announced, emotional intensity rose sharply. *Fear* and *disgust* became the most prevalent emotions, accompanied by a smaller but notable increase in *hate*, marking the most emotionally charged phase of the crisis thus far. These temporal patterns underscore the value of emotion-based indicators for understanding public sentiment during crises. By tracking the relative prevalence of emotional categories, researchers can not only characterize shifts in collective mood but also inform predictive models and communication strategies that respond more effectively to citizens' evolving emotional needs.

V. CONCLUSIONS

SpanishTweetsCOVID-19 provides a structured and comprehensive resource for analyzing pandemic-related discourse, supporting research across disciplines such as information retrieval, social network analysis, misinformation detection, computational social science, public policy, media studies and public health communication. It facilitates both established analytical approaches, such as sentiment analysis and topic modelling, and emerging lines of research, including the long-term effects of crisis communication strategies on public trust and behavior. The collection is designed for accessibility and interoperability, offering detailed documentation on data provenance, preprocessing, and collection methods to promote transparency and reproducibility. Given the still ongoing relevance of pandemic-related research and crisis communication studies, this collection is expected to remain a valuable reference for interdisciplinary investigations.

Despite its contributions, the collection presents some limitations. First, like other social media-based collections, it is inherently constrained to the platform's user base and dynamics, excluding interactions on other social media platforms or in offline contexts. Although *Twitter/X* served as the second major information channel during the pandemic (used by 52% of Argentinians in the most populated provinces, compared to 19% for *WhatsApp*, 16% for *Facebook*, and 13% for *Instagram*, according to Álvarez-Nobell et al. [11]) our collection does not account for cross-platform interactions or wider communication flows. Second, as with any keyword-based collection, biases may arise from term selection and platform-level filtering, which can amplify certain narratives while underrepresenting others. Also, the keyword selection may have not fully captured evolving terminology and shifts in discourse.

Future work could extend this collection by incorporating multi-platform data from *Facebook*, *Instagram*, and traditional media sources, providing a more comprehensive view of pandemic-related discourse. Additionally, the collection could be complemented by *GobBsAsTweets*¹⁰, which compiles the official *Twitter/X* activity of local governments in Buenos Aires Province (Argentina) from 2009 to 2023. This collection focuses exclusively on government–citizen interactions, offering a valuable counterpart for analyzing institutional communication and public engagement dynamics.

Usability could be enhanced by developing interactive tools for data exploration, APIs for easier querying, and predefined splits for benchmarking tasks like crisis response modelling or community modelling. These improvements will strengthen the collection's role as an open, reproducible foundation for understanding how Spanish-speaking communities experienced, communicated, and responded to the COVID-19 crisis.

ETHICAL STATEMENT

Data collection was conducted in full compliance with *Twitter/X*'s Terms of Service and Developer Agreement, using

¹⁰*GobBsAsTweets*: *Twitter* Dataset of Local Governments in Buenos Aires (Argentina): <https://data.mendeley.com/datasets/3fszgrvm2r/2>

only publicly available content retrieved via the official API. To respect user privacy and adhere to platform data-sharing policies, the collection is distributed in the form of *tweet IDs* and preprocessed metadata rather than raw textual content. This design enables researchers to rehydrate the data when necessary or conduct analyses directly on the preprocessed metadata, thereby maintaining compliance with privacy standards and platform policies.

Although tweets are publicly shared, large-scale aggregation introduces ethical challenges related to privacy and anonymity. While *Twitter/X* users consent to public visibility, they may not anticipate that their activity could be analyzed for research purposes. The collection does not include personally identifiable information beyond standard metadata; however, usernames may still indirectly reveal identity. Researchers are encouraged to avoid any attempt at de-anonymization and to prioritize aggregate-level analyses over individual-level reporting. Transparent communication about research objectives and careful data handling are essential for maintaining ethical integrity.

Representation and bias also deserve attention. The collection focuses on Latin American Spanish-language tweets, particularly from Argentina, and therefore does not equally represent all Spanish-speaking communities. Social media activity varies across demographic groups, potentially leading to uneven representation. Moreover, the inclusion of both official and user-generated content means that misinformation, speculation, and polarized opinions may be present. Researchers should account for these factors, acknowledge inherent biases, and implement appropriate mitigation strategies when analyzing or interpreting results. Finally, while this collection is intended for scientific research, potential misuse (such as profiling individuals, amplifying misinformation, or misrepresenting social behavior) remains a risk.

To promote responsible use, researchers should follow institutional ethical guidelines, comply with *Twitter/X*'s evolving data policies, and ensure transparent reporting of findings. Proper citation, attention to user privacy, and adherence to principles of fairness and accountability are strongly encouraged. By following these practices, the research community can responsibly leverage the *SpanishTweetsCOVID-19* collection to advance understanding of crisis communication while minimizing ethical risks.

REFERENCES

- [1] Q. Hu and N. Kapucu, "Information communication technology utilization for effective emergency management networks," *Public Management Review*, vol. 18, no. 3, pp. 323–348, 2016.
- [2] S. C. Kim and K. H. Hawkins, "The psychology of social media communication in influencing prevention intentions during the 2019 us measles outbreak," *Computers in Human Behavior*, vol. 111, p. 106428, 2020.
- [3] G. Lappas, A. Triantafyllidou, and A. Kani, "Harnessing the power of dialogue: examining the impact of facebook content on citizens' engagement," *Local Government Studies*, vol. 48, no. 1, pp. 87–106, 2022.

- [4] G. Lee and Y. H. Kwak, "An open government maturity model for social media-based public engagement," *Government information quarterly*, vol. 29, no. 4, pp. 492–503, 2012.
- [5] P. Palttala, C. Boano, R. Lund, and M. Vos, "Communication gaps in disaster management: Perceptions by experts from governmental and non-governmental organizations," *Journal of contingencies and crisis management*, vol. 20, no. 1, pp. 2–12, 2012.
- [6] J. Zarocostas, "How to fight an infodemic," *The lancet*, vol. 395, p. 676, 2020.
- [7] C. Hevia, P. A. Neumeyer *et al.*, "A perfect storm: Covid-19 in emerging economies," *COVID-19 in developing economies*, vol. 1, no. 1, pp. 25–37, 2020.
- [8] E. Bonsón, D. Perea, and M. Bednárová, "Twitter as a tool for citizen engagement: An empirical study of the andalusian municipalities," *Government information quarterly*, vol. 36, no. 3, pp. 480–489, 2019.
- [9] W. B. Gudykunst and Y. Y. Kim, *Communicating with strangers: An approach to intercultural communication*. McGraw-Hill New York, 1992, vol. 19.
- [10] Y.-H. C. Huang, F. Wu, and Y. Cheng, "Crisis communication in context: Cultural and political influences underpinning chinese public relations practice," *Public Relations Review*, vol. 42, no. 1, pp. 201–213, 2016.
- [11] A. Álvarez-Nobell, A. Oliveira, A. Athaydes, and B. Barroso, "Strategic communication and political ideologies in south america. covid-19 crisis management in the cases of the populist governments of argentina and brazil," *International Journal of Strategic Communication*, vol. 16, no. 3, pp. 403–425, 2022.
- [12] M. del Mar Gálvez-Rodríguez, A. Sáez-Martín, M. García-Tabuyo, and C. Caba-Pérez, "Exploring dialogic strategies in social media for fostering citizens' interactions with latin american local governments," *Public relations review*, no. 2, pp. 265–276, 2018.
- [13] J. Yeo, H. Li, Y. A. Shin, and B. Haupt, "Cultural approach to crisis management," in *Global encyclopedia of public administration, public policy, and governance*. Springer, 2023, pp. 2833–2836.
- [14] N. A. Gemelli, "Management of covid-19 outbreak in argentina: the beginning," *DMPHP*, vol. 14, no. 6, pp. 815–817, 2020.
- [15] A. Tommasel, A. Diaz-Pace, D. Godoy, and J. M. Rodriguez, "Tracking the evolution of crisis processes and mental health on social media during the covid-19 pandemic," *Behaviour & Information Technology*, vol. 41, no. 16, pp. 3450–3469, 2022.
- [16] C. Reuter, A. L. Hughes, and M.-A. Kaufhold, "Social media in crisis management: An evaluation and analysis of crisis informatics research," *Int J Hum-Comput Interact*, vol. 34, no. 4, pp. 280–294, 2018.
- [17] N. Aguilar-Gallegos, L. E. Romero-García, E. G. Martínez-González, E. I. García-Sánchez, and J. Aguilar-Ávila, "Dataset on dynamics of coronavirus on twitter," *Data in brief*, vol. 30, p. 105684, 2020.
- [18] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, and S. Dietze, "Tweetscov19-a knowledge base of semantically annotated tweets about the covid-19 pandemic," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2991–2998.
- [19] C. E. Lopez and C. Gallemore, "An augmented multilingual twitter dataset for studying the covid-19 infodemic," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 102, 2021.
- [20] E. Chen, K. Lerman, E. Ferrara *et al.*, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," *JMIR public health and surveillance*, no. 2, p. e19273, 2020.
- [21] U. Qazi, M. Imran, and F. Ofli, "Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.
- [22] R. Y. Martínez, G. Blanco, and A. Lourenço, "Spanish corpora of tweets about covid-19 vaccination for automatic stance detection," *Information Processing & Management*, vol. 60, no. 3, p. 103294, 2023.
- [23] D. Catalan-Matamoros, I. Prieto-Sanchez, and A. Langbecker, "Crisis communication during covid-19: English, french, portuguese, and spanish discourse of astrazeneca vaccine and omicron variant on social media," *Vaccines*, vol. 11, no. 6, p. 1100, 2023.
- [24] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large arabic twitter dataset on covid-19," *arXiv:2004.04315*, 2020.
- [25] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "Arcov-19: The first arabic covid-19 twitter dataset with propagation networks," *arXiv:2004.05861*, 2020.
- [26] S. Yum, "Who plays an important role for information networks of covid-19 in latin america?" *Disaster Med Public Health Prep*, vol. 18, p. e266, 2024.
- [27] A. Amado, "Citizen's participation on social media against state-sponsored disinformation during the pandemic in argentina 1," in *State-Sponsored Disinformation Around the Globe*. Routledge, 2024, pp. 157–173.
- [28] C. Baumann and L. Denardi, "Covid-19 and digital anti-sinophobia in argentina," *New Diversities*, vol. 24, no. 1, pp. 67–89, 2022.
- [29] F. Arias, A. Guerra-Adames, M. Zambrano, E. Quintero-Guerra, and N. Tejedor-Flores, "Analyzing spanish-language public sentiment in the context of a pandemic and social unrest: The panama case," *Int. J. Environ. Res. Public Health*, vol. 19, no. 16, p. 10328, 2022.
- [30] J. Cabezas, D. Moctezuma, A. Fernández-Isabel, and I. Martin de Diego, "Detecting emotional evolution on twitter during the covid-19 pandemic using text analysis," *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, p. 6981, 2021.
- [31] F. Demonte, D. P. Bruno, L. S. Lozano, M. F. Mena, and A. M. Pereira, "Public conversation on vaccines during the covid-19 pandemic in argentina, 2021-2022," *Salud Colectiva*, vol. 20, pp. e4580–e4580, 2024.

- [32] A. Córdoba-Cabús, M. García-Borrego, and Y. Ceballos, "Sentiment analysis toward the covid-19 vaccine in the main latin american media on twitter: The cases of argentina, chile, colombia, mexico, and peru," *Vaccines*, vol. 11, no. 10, p. 1592, 2023.
- [33] W. Ceron, G. Gruszynski Sanseverino, M.-F. de Lima-Santos, and M. G. Quiles, "Covid-19 fake news diffusion across latin america," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 47, 2021.
- [34] E. M. Heredia, "Prison and covid-19 in argentina through the eyes of twitter users," *Interface: A Journal on Social Movements*, vol. 13, no. 2, 2021.
- [35] I. Guy, *People Recommendation on Social Media*. Springer International Publishing, 2018, pp. 570–623.
- [36] F. Pierri, C. Piccardi, and S. Ceri, "Topology comparison of twitter diffusion networks effectively reveals misleading information," *Scientific reports*, vol. 10, no. 1, p. 1372, 2020.
- [37] —, "A multi-layer approach to disinformation detection in us and italian news spreading on twitter," *EPJ Data Science*, vol. 9, no. 1, p. 35, 2020.
- [38] P. Cogan, M. Andrews, M. Bradonjic, W. S. Kennedy, A. Sala, and G. Tucci, "Reconstruction and analysis of twitter conversation graphs," in *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*. New York, NY, USA: ACM, 2012, pp. 25–31.
- [39] J. Sanz-Cruzado and P. Castells, "Enhancing structural diversity in social networks by recommending weak ties," in *Proceedings of the 12th ACM conference on recommender systems*. New York, NY, USA: ACM, 2018, pp. 233–241.
- [40] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 4647–4657.
- [41] —, "Lexicons on demand: Neural word embeddings for large-scale text analysis," in *IJCAI*, 2017, pp. 4836–4840.
- [42] W. G. Parrott, *Emotions in social psychology: Essential readings*. psychology press, 2001.
- [43] R. Perczek, C. S. Carver, A. A. Price, and C. Pozo-Kaderman, "Coping, mood, and aspects of personality in spanish translation and evidence of convergence with english versions," *J. Pers. Assess.*, pp. 63–87, 2000.
- [44] T. E. Drabek, *Human system responses to disaster: An inventory of sociological findings*. New York, NY, USA: Springer Science & Business Media, 2012.
- [45] B. Richardson, "The phases of disaster as a relationship between structure and meaning: A narrative analysis of the 1947 texas city explosion," *International journal of mass emergencies and disasters*, vol. 23, 2005.
- [46] D. J. DeWolfe, *Training manual for mental health and human service workers in major disasters*. US Department of Health and Human Services, Substance Abuse and Mental ..., 2000.
- [47] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *International AAAI Conference on Weblogs and Social Media (ICWSM)*, no. CONF, 2014.
- [48] M. su Park, "Exploring healthcare opportunities in online social networks: Depressive moods of users captured in twitter," in *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, 2012.
- [49] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, 2017, pp. 277–280.
- [50] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, July 2013.
- [51] O. Gruebner, S. R. Lowe, M. Sykora, K. Shankardass, S. Subramanian, and S. Galea, "Spatio-temporal distribution of negative emotions in new york city after a natural disaster as seen in social media," *Int J Environ Res Public Health*, vol. 15, no. 10, p. 2275, 2018.
- [52] O. Gruebner, S. R. Lowe, M. Sykora, K. Shankardass, S. V. Subramanian, and S. Galea, "A novel surveillance approach for disaster mental health," *PLOS ONE*, vol. 12, no. 7, pp. 1–15, 07 2017.
- [53] J. C. de Albornoz, L. Plaza, and P. Gervás, "SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association, May 2012, pp. 3562–3567.
- [54] J. C. de Albornoz, I. Chugur, and E. Amigó, "Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation," in *CLEF (Online Working Notes/Labs/Workshop)*, P. Forner, J. Karlgren, and C. Womser-Hacker, Eds., 2012.
- [55] M. B. Arnold, *Emotion and personality. Vol. I. Psychological aspects*. Columbia University Press, 1960.
- [56] R. Plutchik, *Emotion: Theory, research, and experience*. NewYork: Academic, 1980, ch. A general Psychoevolutionary Theory of Emotion, pp. 3–33.
- [57] A. Tommasel, A. Diaz-Pace, J. M. Rodriguez, and D. Godoy, "Forecasting mental health and emotions based on social media expressions during the covid-19 pandemic," *IDD*, vol. 49, no. 3, pp. 259–268, 2021.
- [58] A. Tommasel, "A framework for characterising the hows and whens of social media communication during crises," *Journal of Information Science*, 2025.
- [59] L. M. Aiello, D. Quercia, K. Zhou, M. Constantinides, S. Šćepanović, and S. Joglekar, "How epidemic psychology works on social media: Evolution of responses to the covid-19 pandemic," *arXiv*, 2020.
- [60] D. M. Neal, "Reconsidering the phases of disaster," *International Journal of Mass Emergencies & Disasters*, vol. 15, no. 2, pp. 239–264, 1997.