Sentiment and Linguistic Analysis of Epidemic Outbreak Data from Official and Alternative Sources

ARTICLE HISTORY
Received 8 May 2025
Accepted 23 July 2025
Published 6 January 2026

Karina Ordoñez Guerrero
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
karina.ordonez2016@uteq.edu.ec
ORCID: 0009-0009-2507-0519

José Cordero Bazurto
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
jcorderob@uteq.edu.ec
ORCID: 0009-0001-2961-6736

Geovanny Brito Casanova
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
gbritoc@uteq.edu.ec
ORCID: 0000-0002-7715-7706

Eduardo Samaniego Mena
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
esamaniego@uteq.edu.ec
ORCID: 0000-0002-6196-2014

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol XIII, Issue 1, January 2026

Sentiment and Linguistic Analysis of Epidemic Outbreak Data from Official and Alternative Sources

Karina Ordoñez Guerrero
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
karina.ordonez2016@uteq.edu.ec

José Cordero Bazurto
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
jcorderob@uteq.edu.ec

Geovanny Brito Casanova
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
gbritoc@uteq.edu.ec

Eduardo Samaniego Mena
Universidad Técnica Estatal de Quevedo
Facultad de Posgrado
Quevedo, Ecuador
esamaniego@uteq.edu.ec

***Abstract*—** Information on epidemic outbreaks is a key input for health surveillance, as it allows for the assessment of the spread and associated social perception. This study examines emotional and linguistic patterns in narratives disseminated by international organizations (WHO, UN, CDC) and digital platforms (Google News and Reddit) over a three-month period. The KDD process was applied in R Studio (selection, preprocessing, transformation, modeling, and evaluation), using Bing and NRC lexicons and a supervised Naive Bayes model to enhance the detection of emotional nuances. A total of 12,340 texts (3,100 from official sources, 4,240 from Google News, and 5,000 from Reddit) were analyzed using standardized queries in English: pandemic, confinement, epidemic, and HMPV. Official sources showed a greater presence of positive emotions linked to cooperation and security; Google News concentrated negative narratives with terms such as risk and dangerous; Reddit combined fear and sadness with appearances of hope. The analysis included t-tests and ANOVA with 95% confidence intervals. The work is exploratory and preliminary in nature and suggests that surveillance systems should integrate the monitoring of social networks and digital media, along with public policy measures to improve communication in health crisis situations.

***Keywords*—** *epidemic outbreaks, sentiment analysis, text mining, epidemiological surveillance, public communication.*

I. Introduction

The analysis of information on epidemic outbreaks is a key resource for public health, especially in a global scenario where diseases have the capacity to spread rapidly between countries and continents [1]. In addition to the epidemiological course, a practical problem is how narratives influence public confidence and the adoption of measures (e.g., adherence to recommendations or acceptance of vaccines). The heterogeneity of the data poses challenges related to its reliability, consistency, and relevance in responding to immediate needs [2]. Identifying both the strengths and limitations of these sources helps to refine surveillance systems and guide communication strategies during health emergencies [3].

Official sources, represented by the World Health Organization (WHO), the United Nations (UN), and the Centers for Disease Control and Prevention (CDC), were selected due to their recognition in providing validated and structured information used by these institutions apply rigorous protocols in the collection and dissemination of data; however, their immediacy may be limited by administrative processes that delay real-time updates [5]. In contrast, digital platforms such as Google News and Reddit function as alternative sources that allow access to more rapidly disseminated narratives and spontaneously expressed collective perceptions. The use of these sources carries risks, such as the circulation of incomplete or biased information [6].

The choice of Google News and Reddit was based on criteria of coverage and narrative diversity. Google News, as a global news aggregator, compiles publications from various international media outlets, while Reddit focuses on discussions in thematic communities such as r/epidemiology and r/health. Standardized queries with English keywords (pandemic, confinement, epidemic, HMPV) were applied to both platforms, ensuring consistency and facilitating comparisons between sources. However, it is recognized that the inclusion of these platforms introduces biases inherent to the nature of the content: in Google News, due to the media's focus on capturing attention, and in Reddit, due to the influence of communities with particular interests.

The research uses the KDD (Knowledge Discovery in Databases) process, structured in phases of selection, preprocessing, transformation, modeling, and evaluation. In addition to sentiment lexicons (Bing and NRC), a supervised text classification model was incorporated to expand the

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

detection of emotional nuances and evaluate differences between sources. Sentiment analysis provides useful signals for public health by allowing the detection of social alarm peaks, monitoring changes in tone, and prioritizing risk messages.

The main purpose of this study is to compare emotional and linguistic patterns in narratives of epidemic outbreaks from official and alternative sources, analyzing variables such as speed of dissemination, geographical coverage, and emotional charge. The findings describe how risks and expectations are communicated in different settings and offer applicable inputs for strengthening surveillance systems and designing public policy recommendations aimed at improving global health communication.

II. RELATED WORK

The collection and analysis of information on epidemic outbreaks from various sources has enabled advances in the detection, monitoring, and management of diseases globally. Access to official sources, news, and social media platforms has transformed the way data on disease spread is collected, enabling a greater diversity of analytical approaches [11]. Below, we detail how these sources have been used in previous research, highlighting the application of data science to optimize results and overcome challenges inherent in managing large volumes of information.

Official sources, such as the WHO, the UN, and the CDC, have proven to be fundamental tools in consolidating epidemiological data based on formal reports. For example, [12] describes the use of visual tools to explore patterns of epidemic spread in spatial and temporal dimensions, using data from reliable sources to assess the dynamics of diseases such as COVID-19. This approach has proven functional in modeling complex scenarios that require precision in the representation of outbreaks.

At the same time, epidemic propagation models have been analyzed using two-layer networks, where physical and virtual connections allow the spread of information and diseases to be simulated [13]. This approach demonstrates how official data can be integrated with complex simulations to assess the impact of connectivity on the spread of diseases.

In [14], it is documented how text analysis using advanced natural language processing techniques has made it possible to identify trends in the early stages of epidemic outbreaks, thereby improving the ability of official systems to adapt to changing scenarios.

Official sources and alternative sources, such as social media, play a fundamental role in the collection and analysis of data on epidemic outbreaks. In [15], the authors evaluate how the spread of preventive information, whether positive or negative, on multiplex networks can influence public perception and responses to outbreaks. These findings highlight the importance of considering information dynamics in the formulation of health surveillance and response strategies.

The comparative analysis between official and alternative sources has also been addressed in [16], where the probabilities of extreme epidemics occurring were analyzed using historical and current databases. This study highlights how the integration of multiple sources improves the ability to predict and respond to large-scale events.

In recent years, social media has emerged as an alternative source for detecting epidemic outbreaks. For example, [17] analyzes how the spread of information on digital platforms can serve as an early indicator of health events. Their research highlights the use of optimization algorithms to identify critical nodes in these networks, showing how alternative sources complement the limitations of official sources by capturing collective perceptions and behavior patterns in real time. Another study, described in [18], investigated the theoretical limits of inference in epidemic spread through statistical mechanics methods applied to social networks and graphical models. Their work highlights how user-generated data can provide additional insights when integrated with traditional models.

Reddit, in particular, has proven to be an effective source for detecting emerging trends in public perception. The analysis of posts on subreddits such as r/epidemiology has been documented in [19], indicating that they used distributed control models to assess how decentralized networks can influence the spread of information and diseases.

The integration of data science has been fundamental in addressing the limitations inherent in the heterogeneity of information sources. For example, in [20], they developed a model to classify and analyze epidemic information extracted from social networks, achieving high levels of classification accuracy and generating concise summaries of posts.

Additionally, the work in [21] employed advanced methods to infer epidemic trajectories from partial observations, using Bayesian techniques to improve the estimation of dynamic parameters. Such studies show how data science-based models can address challenges related to uncertainty and variability in data.

III. METHODOLOGY

The methodology proposed in this study was designed to analyze information on epidemic outbreaks from official and alternative sources, incorporating text mining, statistical analysis, and visualization techniques. A systematic process was adopted that included the stages of identification, collection, cleaning, analysis, and comparison of data, with the aim of ensuring the accuracy of the results obtained.

A. Study Design

The study is based on a comparative approach that combines automated data consumption techniques and advanced analysis tools to examine characteristics and differences between information sources. These include official platforms, such as the WHO, UN, and CDC websites, along with alternative sources such as Google News and social media, specifically Reddit. The keywords pandemic, confinement, epidemic, and HMPV were applied uniformly across all sources, with the aim of exploring aspects such as the speed of data dissemination, geographic coverage, and the predominant emotions in the reported narratives.

B. Population and Sample

The research worked with a total of 12,340 texts, distributed across 3,100 records from the WHO, UN, and CDC, 4,240 articles from Google News, and 5,000 posts on

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

Reddit. This sample was obtained over a three-month period with periodic collections that facilitated the observation of temporal variations.

1) Official Sources:

These include globally recognized organizations such as the World Health Organization (WHO), the United Nations (UN), and the Centers for Disease Control and Prevention (CDC). These entities were selected due to their ability to generate reliable reports, real-time updates, and detailed analyses of disease spread. The information provided by these sources includes statistics, health alerts, and global reports.

2) Alternative Sources:

This category included Google News and the social network Reddit. Google News was used to access information of interest published by international news media through specific queries related to epidemic outbreaks. Reddit, for its part, offers an organized structure in thematic subreddits such as r/epidemiology, r/health, and r/worldnews, which allow for the identification of publications with technical content and social perceptions about health events.

C. Data collection tool

The tool designed for this research was intended to capture structured information about epidemic outbreaks. Uniform queries using the keywords *pandemic, confinement, epidemic*, and *HMPV* were used in all selected sources (official and alternative). The analysis was performed in R Studio with libraries such as *rvest, httr, jsonlite, tidytext, and ggplot2*, which facilitated extraction, processing, and visualization. Automated queries were configured to collect information directly from available APIs and portals, which allowed for standardization of search criteria and ensured traceability. On Reddit, access was provided through the Reddit API with OAuth2 authentication and JSON requests managed with httr/jsonlite to search endpoints and each subreddit; quotas and usage policies were respected and no scraping was used. For Google News, compatible public feeds (RSS/JSON format) accessible from official aggregator links were used.

D. Collection process

The collection was carried out over a period of three months at regular intervals, allowing for the capture of temporal variations. For Google News, queries were defined using English-language keywords, filtering results into health and science categories. On Reddit, the same keywords were applied to specific subreddits (*r/epidemiology, r/health, r/worldnews*), with filters by language and removal of URLs, emojis, and duplicate content.

It is recognized that these sources introduce biases: in the case of Google News, due to the media's tendency toward impactful narratives, and on Reddit, due to the influence of communities with particular interests. Technical limitations: changes in API endpoints or quotas, intermittent availability of feeds, coverage mainly in English that may exclude local nuances, and lack of access to content that was deleted, private, or moderated before capture.

The collected data was organized into homogeneous structures, and random checks were applied to verify consistency with the original sources. This validation ensured that the records were complete and consistent with the study objectives.

E. Data Analysis

Data analysis was carried out following the five phases of the Knowledge Discovery in Databases (KDD) process, summarized in the diagram in Figure 1 (Source selection → Data collection → Preprocessing → Transformation → Modeling & analysis → Evaluation & visualization). This procedure allowed for the systematic organization of the extraction, cleaning, transformation, modeling, and evaluation of information from official and alternative sources. The implementation was carried out in R Studio using libraries such as rvest, httr, jsonlite, dplyr, tidytext, ggplot2, wordcloud2, stringr, gridExtra, tidyr, and textdata, which facilitated the manipulation, processing, and visualization of textual records.

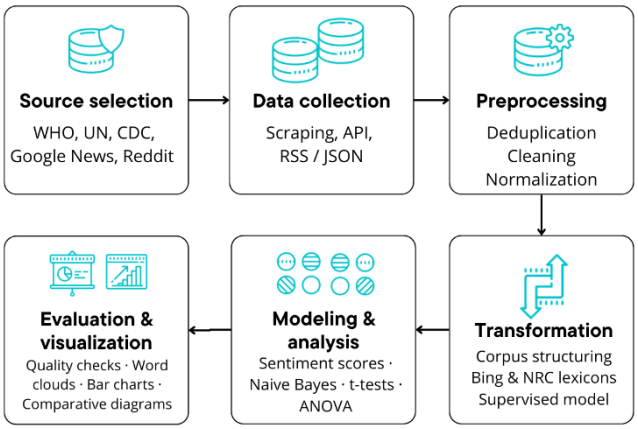

Fig. 1. KDD pipeline used in the study

1) Data Selection:

Reliable sources were selected for the research. Official platforms included the WHO, UN, and CDC, recognized for providing up-to-date epidemiological data. Google News and Reddit were considered as alternative sources. In Google News, news related to specific keywords was analyzed, while in Reddit, thematic subreddits such as r/epidemiology, r/health, and r/worldnews were explored. These platforms provided information on public perceptions and narratives related to epidemic outbreaks.

2) Data Preprocessing:

Cleaning included removing duplicates, normalizing dates, and correcting typos. Texts were converted to lowercase and filters were applied to exclude incomplete entries. The dplyr and stringr libraries were used, ensuring that the final database was aligned with the objectives of the analysis.

3) Data Transformation:

The cleaned records were organized into structures compatible with textual and visual analysis. Tidyr was used to structure the corpus and textdata to associate emotions with the narratives. In addition to the Bing and NRC lexicons, a supervised emotional classification model with cross-validation was trained, which allowed for the detection of nuances not covered in basic dictionaries. This integration

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

facilitated the identification of mixed emotions in the discourses.

4) Modeling and Analysis:

Narrative patterns were represented using ggplot2 and wordcloud2, generating word clouds, bar charts, and scatter plots. All visualizations were produced in English (titles, axes, and legends), with a minimum resolution of 300 dpi and font size ⩾ 10 pt to ensure legibility. Figure 2 includes a color bar with an explicit scale: –1 = negative emotion, 0 = neutral, +1 = positive emotion; the ends of the gradient correspond to those limits. The analysis included the calculation of average sentiment scores per source, as well as the application of Student's t-tests and ANOVA to contrast differences between groups, with 95% confidence intervals.

5) Results Evaluation

The evaluation of the collected and processed data was carried out through a comparative analysis of official and alternative sources, focusing on aspects such as the accuracy, consistency, and representativeness of the information. This process made it possible to verify the quality of the data and ensure its alignment with the research objectives. To this end, several methodological aspects were evaluated:

- **Accuracy of information:** Cross-checks were performed with the original reports from each platform, confirming that the data extracted through automated queries corresponded to verifiable narratives. This validation was useful for discarding duplicate records or those outside the defined time range.
- **Consistency between sources:** Emotional and thematic patterns were analyzed to determine the correspondence of results with the initial queries (pandemic, confinement, epidemic, HMPV). The contrast between narratives showed differences linked to the biases of each source: a tendency toward high-impact content on Google News and a predominance of community perspectives on Reddit.
- **Data representativeness:** The diversity of terms and emotions was verified, confirming that the three sets offered a balanced sample of institutional, media, and social narratives. An analysis of geographical and temporal coverage was included, which ensured that the data reflected variations at different stages of the outbreaks.

In addition, calculations of average sentiment scores and contrast tests were added to the textual metrics. Student's t-test and ANOVA were applied, yielding statistically significant differences between sources in the dimensions of trust, fear, and sadness. Confidence intervals of 95% were reported, reinforcing the validity of the comparisons.

The figures were generated in English to maintain consistency. Word clouds, bar charts, and scatter plots were used. Each figure included complete descriptions of data source, time range, and color meaning. Color gradients differentiated emotional intensity: dark tones indicated negative emotions, while light tones represented positive or neutral emotions. This visual coding allowed for quick and accurate interpretation of emotional variations.

IV. RESULTS

The data analysis, carried out in January 2025, focused on the processing and evaluation of textual information extracted from three main sources: official and alternative platforms. This approach combined advanced text mining tools, sentiment analysis, and data visualization using word clouds and bar charts, allowing for a comprehensive interpretation of the emotional and linguistic narratives present in each source. To increase rigor, the emotional analysis was not only performed with Bing and NRC lexicons, but also applied a Naive Bayes classifier trained with manual annotations, which allowed for the identification of nuances beyond lexical detection. In addition, sentiment scores were normalized to a range of −1 to +1, and t-tests and ANOVA with 95% confidence intervals were applied, confirming differences between sources.

The visualizations generated, such as Figure 2, used a color gradient bar designed to represent emotions and their intensity. Blue tones were associated with feelings of security and confidence, reflecting positive emotions that promote calm and optimism. In contrast, reddish tones indicated negative emotions such as danger, fear, and risk, allowing us to identify areas of greater alarm or concern in the analyzed texts. This gradient is an effective visual resource for highlighting emotional transitions in the content and has been described in English in all figures to maintain consistency with the language of the processed terms.

Word clouds were used to highlight the most frequent terms in each dataset. In these, the size of the words reflected their prominence within the text, while the intensity of the color represented the average polarity calculated by the model. Warm tones such as orange and red indicated words with negative connotations, while cool, bluish tones highlighted terms associated with more positive or neutral emotions.

These tools made it possible to identify specific patterns and trends in the narratives of each source, helping to structure the results in a comprehensible and verifiable manner. In addition, it was observed that the presentation characteristics of each platform (headlines and thematic hierarchy in aggregators; institutional language and editorial guidelines in official bodies; votes and moderation on Reddit) shape the emotional expression of the content and its reception by the public. Figure 2 is an illustrative example of how colors and shades allow us to capture the emotional range contained in the analyzed texts, facilitating an interpretation based on statistical metrics.

Fig. 2. Color gradient bar

A. Analysis of Data from Official Sources

Official websites provided reliable and structured epidemiological data, designed to communicate messages focused on practical solutions and concrete measures to mitigate the effects of epidemic outbreaks. Sentiment analysis showed a predominance of positive emotions, such as confidence and anticipation, in line with the goals of inspiring

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

calm, promoting international cooperation, and encouraging collaborative action of these institutions. Although negative emotions such as fear and sadness were present, they were addressed in a controlled manner, with an emphasis on empathy and efforts to overcome the challenges associated with health crises.

In addition to lexicons, a supervised classifier for emotions was applied, confirming higher average scores in positive categories. Comparisons between valences were evaluated with Student's t-test, reporting $p < 0.05$ and 95% CI not including 0, which supports the difference observed in this group. It should be noted that the queries used were standardized in English (pandemic, confinement, epidemic, HMPV).

Figure 3 illustrates how official websites prioritize positive emotions, such as trust, over negative emotions, which, although they exist, are presented with less alarm. This representation reinforces the optimistic nature of their narratives, aimed at generating security and trust in the public.

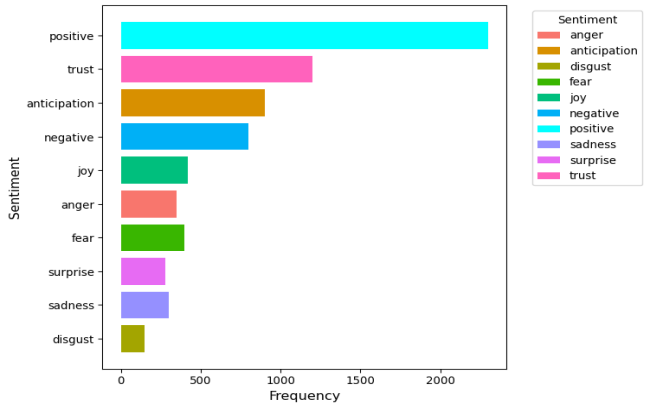

Fig. 3. Distribution of sentiments on official pages

On the other hand, the word cloud generated from official publications highlights terms such as secure, support, protection, and emergency, associated with collective action and cooperation between institutions and citizens. Figure 4 shows this cloud: the size of each word corresponds to its frequency of appearance, and the predominance of blue indicates its link to positive emotions such as security and trust.

In addition, the color gradient of the cloud helps to distinguish the emotional charge: more intense blues indicate words linked to positive and recurring emotions; light blues highlight less frequent terms, although with optimistic connotations. Reds, present to a lesser extent, mark terms associated with alert or danger, which makes it easier to recognize nuances in the discourse of official sources.

Fig. 4. Word cloud generated from official sources

B. *Analysis of Data from Google News*

Analysis of data obtained from Google News showed a predominance of negative emotions, such as sadness and fear, associated with the editorial tendency to highlight adverse and high-impact aspects. Anger was also observed, related to perceptions of social frustration about the management of the health crisis and its side effects. Positive emotions, such as confidence and joy, appeared to a lesser extent, indicating less attention to narratives of progress or achievements. This pattern may reinforce the perception of risk in audiences with high exposure to alarming headlines.

Standardized queries in English (pandemic, confinement, epidemic, HMPV) were used in health/science categories. Editorial bias toward high-impact content is recognized, favoring a greater presence of negatively charged terms. Normalized sentiment scores and ANOVA between groups showed a greater negative charge in this source ($p < 0.01$; 95% CI of the difference > 0).

Figure 5 shows how negative emotions are more prevalent in Google News content. This pattern coincides with media narratives focused on severity and urgency.

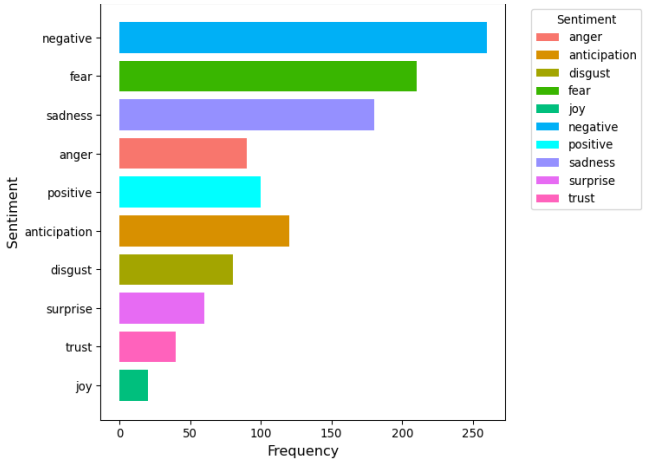

Fig. 5. Distribution of sentiments in Google News

The word cloud generated from these posts (Figure 6) supports this observation, highlighting terms such as fever, dead, confined, stress, outbreak, and dangerous. These words point to risks, restrictions, and unfavorable consequences, resulting in an emotionally charged narrative. Terms such as worst, prisoner, and risk reflect experiences with social and psychological effects, while words in dark red tones, such as dead and dangerous, reinforce the prevailing negative tone. In

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

contrast, hope and relief appear less frequently and with less visual prominence.

The chromatic degradation used in the cloud helps to distinguish the emotions associated with the terms analyzed. Intense reddish tones indicate words linked to negative emotions, while lighter warm tones reflect the low frequency of terms with positive perspectives, such as protection or solidarity. This allows us to see how the intensity of the color and the size of the words convey useful information about the emotional charge and importance of the terms in the context analyzed.

Fig. 6. Word cloud generated from Google News

C. *Analysis of Data from Reddit*

Reddit was identified as a platform where users share experiences, opinions, and emotions related to epidemic outbreaks. The analysis was based on specific queries in subreddits such as r/epidemiology, r/health, and r/worldnews. These communities allow us to observe posts with diverse perspectives on prevailing concerns and emotions related to outbreaks. The same queries were applied in English and filtered by language, removing URLs, emojis, and repetitions to reduce noise. Community bias is recognized due to the rules and culture of each subreddit, which can skew topics or tones.

Sentiment analysis showed that fear was the most recurrent emotion, associated with uncertainty about the impact of outbreaks, mobility restrictions, and perceived risks. Sadness also featured prominently, reflecting the social and emotional effects of crises, such as isolation and human loss. At the same time, emotions of anticipation were detected, linked to expectations about medical advances, government decisions, and possible solutions.

The supervised classifier detected affective mixtures, and Student's t-tests between polarities yielded $p < 0.05$ with a 95% CI not including 0, supporting the predominance of negative emotions over positive ones in this source. Figure 7 shows the distribution of emotions in Reddit posts, with fear and sadness predominating. Positive emotions, such as confidence and joy, appear less frequently, indicating a focus on concerns and challenges.

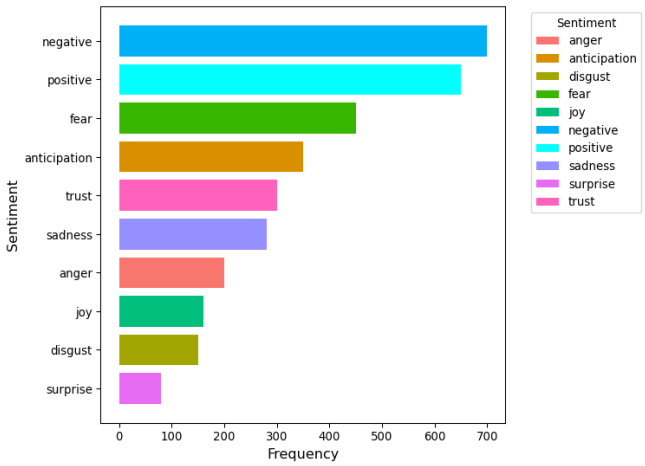

Fig. 7. Distribution of feelings in posts extracted from Reddit

The word cloud generated from the posts (Figure 8) reveals prominent terms such as virus, outbreak, infections, risk, and epidemic. These words reflect concerns about the magnitude of the outbreaks and the severity of their consequences. Terms such as bad and severe reinforce an alarmist tone in the discussions, while positive and secure appear less frequently.

The size of the words in the cloud reflects their frequency in the posts: the largest terms, such as virus and risk, are those that appear most frequently. This visual representation helps to identify the most discussed topics. In addition, the color degradation applied allows the emotional intensity of the terms to be distinguished. Dark tones highlight words associated with negative emotions, such as risk and severe, while warm, light tones, present in positive, show a limited presence of optimistic ideas. This visual approach was complemented by the supervised classifier, confirmed by testing with IC 95%.

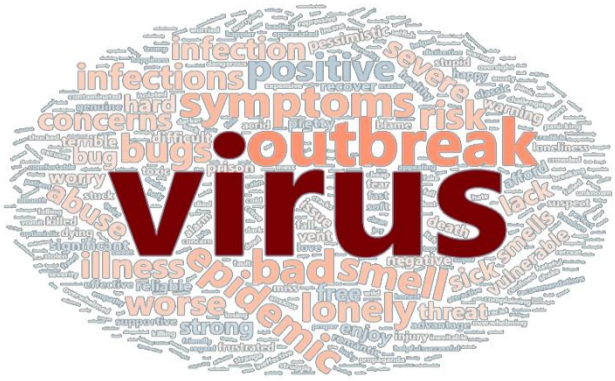

Fig. 8. Word cloud generated from Reddit posts

Reddit, therefore, is identified as a space where collective discussions reflect social concerns and, at the same time, attempts at resilience in the face of epidemic outbreaks. The use of size and colors in the word cloud facilitates an intuitive graphical representation, showing how users emotionally address the challenges associated with the pandemic.

D. *Comparison between official and alternative sources*

The comparative analysis between official and alternative sources highlights marked differences in narratives and

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

emotions about epidemic outbreaks, showing how each group communicates information and how these narratives influence public perception.

Official sources, represented in fuchsia, focused on promoting messages of trust, cooperation, and security. Terms such as secure, websites, official, and organization dominated the narratives, reflecting language structured to generate calm. The size of these words in the cloud illustrates their prominence, while the light tones reinforce the positive emotional charge. These institutions prioritized collaborative action and concrete measures to mitigate the effects of the outbreaks. As shown in Figure 9, these guidelines seek to stabilize public perception and promote international cooperation.

On the other hand, Google News, identified in purple, presented a more alarmist narrative, focusing on the impact of the outbreaks and their risks. Terms such as pandemic, risk, confined, and dangerous emerge as the most prominent, with darker tones intensifying the negative connotation. This approach, reflected in Figure 9, points to content that highlights severity and urgency, with potential effects on risk perception and collective stress.

Reddit, represented in green, showed a more diverse and participatory narrative, with terms of alarm and also of hope. Words such as virus, epidemic, and lock were frequent, while hope and secure reflect attempts to regain optimism. The green tones, with gradations between light and dark, illustrate the mix of emotions, from fear to hope. As shown in Figure 8, Reddit functions as a space for collective expression where emotions and personal experiences coexist.

Figure 9 summarizes these differences in a comparative visual diagram. Official sources stand out for their optimistic approach with light tones and security-oriented terms. Google News prioritizes alarming narratives through dark tones that convey urgency and gravity. Reddit offers a balance between negative and positive terms, with a wide range of emotions and perspectives. This analysis allows us to see how each type of source participates in the construction of perceptions during health crises.

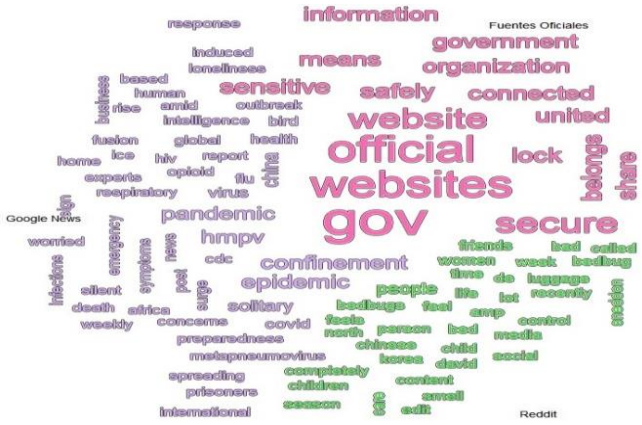

Fig. 9. Comparison between official and alternative sources

V. DISCUSSION

The results obtained in this study show notable differences in the representation and perception of emotions related to epidemic outbreaks on platforms such as official and alternative sources. Beyond the metrics, it can be observed that the editorial framing and dissemination logic of each platform can elevate or attenuate the perception of risk, with practical effects on public attention and willingness to take action.

A. Reddit as a space for social expression

On Reddit, a predominance of fear and sadness was identified, accompanied by anticipation. These emotions respond to the participatory nature of the platform, where users share personal experiences and collective opinions without institutional filters. This bias toward experiential accounts reinforces discussions oriented toward alertness and concern, so Reddit works better as a mood barometer than as a verifiable information channel for immediate decisions.

B. Optimistic narrative in official sources

Official sources were characterized by more controlled communication, focusing on positive feelings such as confidence and anticipation. These institutions seek to convey calm and encourage cooperation through messages designed to avoid panic.[4] and [6] support this trend, indicating that international organizations tend to prioritize narratives that promote security in times of crisis. However, criticism of this strategy was also identified, as it can seem distant from individual experiences, as mentioned in [19] . This disconnect can lead to a perception of institutional coldness in times of uncertainty.

C. Alarmist narrative in Google News

Analysis of Google News publications showed a predominance of negative emotions, such as fear, sadness, and anger. This approach emphasizes the most alarming aspects of the news, aligning with studies such as [2], which indicate that the media often prioritize shocking content to capture the attention of their audiences. Terms such as “confined,” “dangerous,” and “risk” are recurring examples in this narrative. [15], documents how this type of approach can amplify the public's perception of risk, generating knock-on effects. [11] warns that this trend can increase stress in audiences and make it difficult to make informed decisions.

The word cloud generated from Google News reinforces this observation, showing a greater representation of terms related to risks and limitations. Although words such as hope and relief are present, their lower frequency reflects a reduced interest in highlighting progress or positive elements. This highlights the need to balance media narratives with messages that promote resilience and confidence.

D. Impact of technologies and data analysis

The use of advanced technologies such as APIs, text mining, and sentiment analysis was central to this study. Tools such as rvest, httr, ggplot2, textdata, and wordcloud2 facilitated both the analysis and visualization of large volumes of data, allowing for the exploration of emotional and linguistic patterns. The combination of these methods with lexicons such as Bing and NRC made it possible to identify predominant emotions and map how they are reflected in the narratives of different platforms.

In addition, these technologies organized the data and showed how variables such as size, color intensity, and shades in word clouds help to interpret the results visually. For example, the use of gradients made it possible to distinguish the emotions associated with specific terms, which facilitated the detection of differentiated narrative and emotional patterns between official and alternative sources.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026
DOI: 10.5281/zenodo.17288347

VI. CONCLUSIONS

The analysis showed that official sources, such as those managed by the WHO, UN, and CDC, are geared toward conveying confidence and promoting cooperation among audiences. These narratives seek to minimize panic and generate security through structured messages. Although less prevalent, negative emotions also appeared, associated with the communication of risks and challenges, which were handled in a controlled manner to avoid unnecessary alarm.

On the other hand, alternative platforms such as Google News showed a tendency toward alarmist narratives that highlight elements of risk and uncertainty, which can increase the perception of stress in audiences. The predominant presence of terms linked to dangers and restrictions shows an editorial style focused on capturing attention through high-impact headlines.

Greater neutrality was expected on Reddit; however, the analysis revealed a mixture of alarmist accounts with isolated attempts at optimism. Discussions tend to focus on intense concerns and a range of negative emotions, with occasional appearances of hope and resilience

These results should be viewed as preliminary and exploratory. Based on them, practical applications are proposed: real-time monitoring dashboards that integrate media and social media signals, early warnings based on polarity shifts, and segmented communication campaigns that combine institutional messages with readings of the public's emotional state. Together, they offer a comprehensive overview of how the narratives of each source influence collective perception during health crises.

REFERENCES

[1] K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks," *Kurdistan Journal of Applied Research*, pp. 54–65, May 2020, doi: 10.24017/covid.8.

[2] A. Joshi, S. Karimi, R. Sparks, C. Paris, and C. Raina Macintyre, "Survey of Text-based Epidemic Intelligence," *ACM Comput Surv*, vol. 52, no. 6, Nov. 2019, doi: 10.1145/3361141.

[3] K. Sherratt *et al.*, "Characterising information gains and losses when collecting multiple epidemic model outputs," *Journal Epidemics*, vol. 47, Jun. 2024, doi: 10.1016/J.EPIDEM.2024.100765.

[4] J. A. Polonsky *et al.*, "Outbreak analytics: a developing data science for informing the response to emerging pathogens," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2019, doi: 10.1098/rstb.2018.0276.

[5] K. O. Bazilevych *et al.*, "Information system for assessing the informativeness of an epidemic process feature," *System research and information technologies*, vol. 2023, no. 4, pp. 100–112, Dec. 2023, doi: 10.20535/SRIT.2308-8893.2023.4.08.

[6] A. N. Desai *et al.*, "Real-time Epidemic Forecasting: Challenges and Opportunities," *Health Secur*, vol. 17, no. 4, p. 268, Jul. 2019, doi: 10.1089/HS.2019.0022.

[7] J. L. Herrera-Diestra, J. M. Buldú, M. Chavez, and J. H. Martínez, "Using symbolic networks to analyse dynamical properties of disease outbreaks," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, no. 2236, Apr. 2020, doi: 10.48550/arXiv.1911.05646

[8] T. H. Nguyen, M. Fisichella, and K. Rudra, "A Trustworthy Approach to Classify and Analyze Epidemic-Related Information From Microblogs," *IEEE Trans Comput Soc Syst*, 2024, doi: 10.1109/TCSS.2024.3391395.

[9] J. Tolles and T. Luong, "Modeling Epidemics With Compartmental Models," *Journal Jama Network*, vol. 323, no. 24, pp. 2515–2516, Jun. 2020, doi: 10.1001/JAMA.2020.8420.

[10] M. Marani, G. G. Katul, W. K. Pan, and A. J. Parolari, "Intensity and frequency of extreme novel epidemics," *Proceedings of the National Academy of Sciences*, 2021, doi: 10.1073/pnas.2105482118/-/DCSupplemental.

[11] A. Braunstein, L. Budzynski, and M. Mariani, "Statistical mechanics of inference in epidemic spreading," *Phys Rev E*, vol. 108, no. 6, Dec. 2023, doi: 10.1103/PhysRevE.108.064302.

[12] J. Wu, Z. Niu, and X. Liu, "Understanding epidemic spread patterns: a visual analysis approach," *Health Systems*, vol. 13, no. 3, pp. 229–245, Jul. 2024, doi: 10.1080/20476965.2024.2308286.

[13] S. Gracy, P. E. Pare, H. Sandberg, and K. H. Johansson, "Analysis and distributed control of periodic epidemic processes," *IEEE Trans Control Netw Syst*, vol. 8, no. 1, pp. 123–134, Mar. 2021, doi: 10.1109/TCNS.2020.3017717.

[14] K. M. A. Kabir and J. Tanimoto, "Analysis of epidemic outbreaks in two-layer networks with different structures for information spreading and disease diffusion," *Commun Nonlinear Sci Numer Simul*, vol. 72, pp. 565–574, Jun. 2019, doi: 10.1016/J.CNSNS.2019.01.020.

[15] Z. Wang, C. Xia, Z. Chen, and G. Chen, "Epidemic Propagation with Positive and Negative Preventive Information in Multiplex Networks," *IEEE Trans Cybern*, vol. 51, no. 3, pp. 1454–1462, Mar. 2021, doi: 10.1109/TCYB.2019.2960605.

[16] B. Wang, M. Gou, and Y. Han, "Impacts of information propagation on epidemic spread over different migration routes," *Nonlinear Dyn*, vol. 105, no. 4, pp. 3835–3847, Sep. 2021, doi: 10.1007/S11071-021-06791-8/METRICS.

[17] Z. Wang, X. Rui, G. Yuan, J. Cui, and T. Hadzibeganovic, "Endemic information-contagion outbreaks in complex networks with potential spreaders based recurrent-state transmission dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 573, Jul. 2021, doi: 10.1016/J.PHYSA.2021.125907.

[18] S. S. Chikkaraddi and G. R. Smitha, "Epidemic Disease Expert System," *1st IEEE International Conference on Advances in Information Technology, ICAIT 2019 - Proceedings*, pp. 571–576, Jul. 2019, doi: 10.1109/ICAIT47043.2019.8987421.

[19] K. Osadcha, V. Osadchyi, and V. Kruglyk, "The role of information and communication technologies in epidemics: an attempt at analysis," *Ukrainian Journal of Educational Studies and Information Technology*, p., 2020, doi: 10.32919/uesit.2020.01.06.

[20] M. Imanipour, M. Shahmari, | Saeideh, A. Mahkooyeh, A. Ghobadi, and P. Sanjari, "Reflections on health information sources in epidemics in synchrony with the COVID-19 pandemic: A scoping review," *Journal of Nursing Advances in Clinical Sciences*, vol. 1, 2024, doi: 10.32598/JNACS.2401.1005.

[21] S. L. Peng *et al.*, "NLSI: An innovative method to locate epidemic sources on the SEIR propagation model," *An interdisciplinary Journal of NonLinear Science*, vol. 33, no. 8, Aug. 2023, doi: 10.1063/5.0152859.

AUTHORS

Karina Ordoñez Guerrero

Karina Michelle Ordóñez Guerrero is a Systems Engineer and holds a Master's Degree in Data Science from the Technical State University of Quevedo (UTEQ). She has experience in programming, data analysis, and developing technological solutions. She collaborates with teachers on projects involving applied analytics, text mining, and visualization, contributing to experimental design and the creation of scripts in R and Python.

During her pre-professional internship at the University Wellness Unit (UBU), she participated in data collection, processing, and analysis, provided technical support, and helped generate reports to improve institutional processes. At the National Institute of Statistics and Census (INEC), she was involved in the planning and supervision of operational processes focused on the management and updating of cartographic data, standing out for her organization and results-oriented approach.

She is currently a programmer and technical assistant at the Royal Dental Center, where she develops solutions that optimize operational and administrative processes. In addition, she supports startups as a technical assistant, gathering requirements, prototyping, testing, and implementing web and mobile applications, and providing technological support to teams starting projects. She has participated in scientific conferences and training sessions on artificial intelligence, data analysis, and emerging technologies, strengthening her technical profile.

José Cordero Bazurto

José Steven Cordero Bazurto is a Systems Engineer and holds a Master's Degree in Data Science from the Technical State University of Quevedo (UTEQ). He works as a programmer and researcher in technological applications and is affiliated with UTEQ, where he collaborates on innovation projects. He is currently a developer of UTEQ's Academic Management System (SGA), participating in the design, development, and improvement of academic modules, database integration, and information analysis to support institutional decision-making.

He is the author of scientific articles in indexed journals, including a publication in Ciencia Huasteca from the Higher School of Huejutla. His areas of expertise include software development, data analytics, and visualization, with experience in requirements gathering, architecture design, API creation, dashboard construction, and data flow automation. At UTEQ, he has collaborated with academic teams on initiatives aimed at improving processes through web solutions and integrated services.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026

AUTHORS

Geovanny Brito Casanova

Geovanny José Brito Casanova has a degree in Systems Engineering from the Quevedo State Technical University (UTEQ), where he is currently a lecturer at the Faculty of Computer Science and Digital Design. He holds a Master's degree in Development and Operations (DevOps) from the International University of La Rioja (Spain) and a Master's degree in Data Science from UTEQ.

During his academic training, he was recognized for his excellent academic performance within his degree program and faculty, receiving institutional distinctions and being awarded national and international postgraduate scholarships. His academic and professional experience focuses on the development and implementation of technological solutions, particularly in the areas of education, data science and cloud computing. He has collaborated as a reviewer for scientific journals and has participated as a speaker in academic events with national and international reach.. His research work covers topics such as educational software, digital infrastructure, environmental automation and the use of new technologies in educational processes.

He is currently involved in university research projects that focus on data analysis, the development of digital environments and the improvement of educational processes through technology.

Eduardo Samaniego Mena

Eduardo Amable Samaniego Mena holds a degree in Systems Engineering from the Technical State University of Quevedo (UTEQ), a Master's degree in Connectivity and Computer Networks from UTEQ, and a Master's degree in Visual Analytics and Big Data from the International University of La Rioja (UNIR, Spain). He is an undergraduate and graduate professor at UTEQ and a tenured professor. His academic activity revolves around applied research, with an emphasis on computer networks, data analytics, and visualization.

In his research work, he develops solutions based on text mining, machine learning, and statistical analysis, integrates end-to-end data pipelines, and builds analytical dashboards for decision-making. He participates in projects with multidisciplinary teams, supervises degree theses, and publishes results aimed at solving real-world problems. He uses tools such as Python, R, SQL, Power BI, Tableau, and network infrastructure and security technologies, combining good engineering practices with scientific methodology.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2026