Big Data-Driven Social Media Analytics Using a COVID-19 Dataset

Phase 3 Report

Ritwik Raj Saxena   Doondi Ashlesh Tammineedi   Austin Harkins

Abstract

Social media continues to proliferate rapidly and generate substantial volumes of user-generated content. In the field of COVID-19, user-generated data on Twitter can be utilized by policymakers, researchers, and healthcare professionals for seeking to understand public sentiment and behavior in real time. This is imperative to guarantee more informed, timely, and targeted interventions during a public health crisis and to prepare for future pandemics in an increasingly skeptical world with a swiftly growing population. It gives unprecedented opportunities to recognize the behavioral patterns and opinions of social media users with regards to COVID-19. Our study presents a big data-driven approach for analyzing a large dataset of tweets associated with COVID-19. We utilize modern natural language processing techniques and approaches, including but not limited to transformer-based embeddings, clustering, and zero-shot classification, we automatically extract latent topics, sentiments, bot-generated tweets and outliers. The framework further provides actionable insights from the dataset. We aim to demonstrate that large-scale analysis of COVID-19 tweets has the potential to reveal distinctive patterns which are useful for research and formulation of public policy. Our goal is to highlight the potential of big data analytics to transform semi-structured textual data into insightful and actionable information.

1 Introduction

Social media continues to accelerates the quantum shift in the patterns of how individuals communicate, express opinions, and share experiences across the globe. As of 2025, over 5 billion people, more than 60% of the world's population, actively use social media platforms such as X (formerly Twitter), Facebook, Instagram, TikTok, and YouTube [1]. Recent data from October 2025 indicates there are approximately 5.66 billion active social media users, which equates to about 68.7% of the global population [2]. This massive user base contributes to an ever-expanding reservoir of user-generated content which is estimated to exceed several petabytes of new data daily. Such content encompasses text, images, videos, and live streams and encapsulates various perspectives on real-world events in real time. The accessibility and immediacy of social media have thus made it a vital source of information for researchers, policymakers, and businesses. The increasing integration of AI-driven recommendation systems and multimodal analytics further accelerates content creation, reshapes online discourse and amplifies the influence of social media on public opinion, consumer behavior, and crisis communication.

In the context of the COVID-19 pandemic, user-generated data on Twitter has arisen as an analytical resource for gauging public sentiment, behavior, and information dissemination in real time. During the pandemic, millions of users worldwide posted tweets containing opinions, experiences, and reactions to evolving public health policies, vaccination campaigns, and emerging variants and developed an enriched data source. Studies have shown that Twitter activity can reflect public anxiety, misinformation spread, and attitudes toward preventive measures such as mask-wearing and social distancing [3]. During the early months of the pandemic, there was a spike in negative emotions coinciding with lockdown announcements and rising case counts [4]. Such occurrences provide early warning signals for public health authorities. Scrutinizing tweets has the potential to enable researchers to track regional differences in compliance, vaccine hesitancy, and misinformation propagation and inform targeted interventions [5]. Policymakers and healthcare professionals can hence use such data to design evidence-based communication strategies, monitor the effectiveness of health campaigns, and anticipate behavioral responses. In this way, social media can prove to be an indispensable tool for real-time epidemiological insight and crisis management, subsuming in its ambit public health crises [6].

1.1 Problem Statement

There is a significant challenge in extracting actionable insights from these data sources in real time, particularly during public health crises such as COVID-19, despite the rapid proliferation of social media and the vast volumes of user-generated content it produces. Policymakers, researchers, and healthcare professionals face difficulties in understanding public sentiment, behavioral patterns, and misinformation due to the unstructured, high-dimensional, and noisy nature of social media data. Manual probing of enormous numbers of COVID-19 tweets is infeasible at scale and while existing analytics has performed a lot in the arena of sentiment estimation, it shows inadequacies in encapsulating dormant topics, detect bot-generated content, and identify outliers successfully. We therefore believe that there exists the need for a framework that uses modern natural language processing and big data analytics to systematically process large datasets of COVID-19-related tweets and empower timely, evidence-based interventions and informed public policy decisions.

A substantial portion of COVID-19 content on social media is inaccurate and misleading. Studies indicate that nearly one in four COVID-19-related health tweets contains erroneous claims such as false treatment methods and unverified preventive measures [7]. This is termed an infodemic and it amplifies public confusion, reduces compliance with official guidelines, and undermines vaccination campaigns [8]. The huge volume of social media activity makes it impossible for health authorities to manually monitor and correct every misleading post. This challenge motivates the development of automated systems capable of identifying COVID-19-associated misinformation on social media.

Public sentiment and discourse around COVID19 vary significantly across different populations and regions. Research shows that tweet volumes, topics, and expressed concerns correlate with geographic location, racial and ethnic composition, and socio-economic status [9]. Some communities discuss economic and health-related impacts of COVID-19 more frequently but others might show differing levels of awareness and skepticism [10]. Identifying these disparities is important for equitable and inclusive public health responses. Analysis of large-scale social media data enable policymakers to customize and individualize interventions and communication strategies for specific populations to see that public health measures reach and resonate with diverse groups effectively.

Information on platforms like Twitter spreads with remarkable speed and complexity. COVID-19-related discussions precede real-world events and outbreaks. Analytics of certain tweet patterns shows the capability to predict spikes in infection rates days in advance. There is a need for real-time analytical tools in this arena. This capability is singularly vital in fast-moving crises where delays in detection can have severe consequences. The enormity in the number of COVID19-related tweets shows that there is a vast repository of semi-structured data in this domain that can be tapped. Thakur [11] documents in his study on long COVID discussions that more than 1.2 million tweets between May 2020 and January 2023 were dedicated to personal accounts of prolonged symptoms. The sheer number of self-reported experiences reflects both the prevalence of long COVID concerns and the overwhelming amount of raw data researchers had to contend with. In this data lie rich behavioral signals, public opinions, and emergent trends that could inform policy decisions, health communication strategies, and research. Big data-driven natural language processing frameworks offer a solution by converting raw tweets into structured, analyzable information.

The propagation of COVID-19 information is heavily influenced by non-official actors. These actors include bots, influencers, and coordinated campaigns. Research has shown that expert and high-profile user tweets achieve far greater reach than official government communications [12]. However, bot-generated content has the capacity to distort public perception and amplify uninformed, inexact, and distortive accounts [13]. Structural features such as hashtags, mentions, and retweet networks play a significant role in how information spreads. There is a need to detect tweets from such actors and anomalous content patterns for understanding the true drivers of public discourse.

One of the problems linked to analytics of COVID-19 social media data was the dynamic and evolving nature of online language during the pandemic. As Khandelwal et al. [14] emphasize, conventional approaches often fail to capture emergent vocabulary and newly coined terms that gain traction in public discourse. During COVID-19, the lexicon of social media expanded rapidly to include novel hashtags, slang, and technical jargon related to the virus, vaccines, and variants. Mental health conversations became saturated with expressive but unconventional linguistic markers. Many of them were those which standard sentiment dictionaries could not interpret. This dynamic evolution posed a severe risk of blind spots in sentiment monitoring. If analytical systems failed to recognize and adapt to emergent terms, large volumes of relevant discussions would remain unclassified. This would leave decision-makers unaware of critical shifts in public mood. Such problems delayed recognition of growing anxiety, depression, or resistance to public health measures. This would undermine crisis management efforts.

The multifaceted nature of public fears during the pandemic also exemplifies the problem. Sayed et al. [15] demonstrate how conversations intertwined concerns about health with fears of economic collapse. Such concerns were often expressed simultaneously in online discussions. This dichotomy reveals a complication. The complication emphasizes that pandemic sentiment cannot be reduced to a single dimension. People were not only worried about infection and death but also about job losses, business closures, and social instability. A big data analysis system that ignores one side of this complex reality risks misrepresenting public priorities. The difficulty of balancing these intertwined concerns shows why simplistic approaches to analysis of COVID-19 social media data are inadequate.

Societal shifts such as remote work and online education created additional sentiment challenges. Vohra and Garg [16] point out that the sudden global transition to remote working generated polarized discussions online. Some groups celebrated flexibility while others voiced stress and dissatisfaction. These contrasting views complicate the task of deriving actionable insights from social media because policymakers and organizations could misinterpret public attitudes if data analysis tools fail to account for heterogeneity. Waheeb et al. [17] and Sosun et al. [18] highlight the immense disruption in education. In this regard, propaganda, confusion, and distrust about online learning flourished. The spread of fake news regarding e-learning compounded parents' and students' anxieties. This created an environment where negative sentiment toward digital education threatened the continuity of academic systems. Such sentiment must be taken into account to build trust in essential digital infrastructures and ensure educational recovery during and after the pandemic.

1.2 Motivation

The relentless growth of social media, particularly during the COVID-19 pandemic, has created an unprecedented opportunity as well as a critical challenge for public health and policy. However, this wealth of semi-structured data is plagued by inherent complexity: it is noisy, high-dimensional, and severely compromised by the dumping of huge amounts of misinformation and bot-generated content, which existing analytics struggle to isolate and correct. The motivation for this study is therefore to develop a novel, big data-driven framework that utilizes modern natural language processing techniques to systematically process this corpus of COVID-19 tweets. Our goal is to engage in sentiment assessment, identify behavioral outliers, and detect coordinated information campaigns.

1.3 Contributions

1. We perform topic modeling on COVID-19 tweets in order to discover the underlying thematic structure and key subjects being discussed.

2. Sentiment estimation is conducted on a dataset of COVID-19 tweets to classify the emotional tone and attitude expressed in each tweet.

3. Tweets that were posted by bots are segregated to identify and quarantine content generated by automated accounts.

4. Outlier detection is also conducted to identify tweets that deviate significantly from the norm. Anomalous tweets include potential spam and misinformation posted by inauthentic accounts.

2 Related Works

Previous research has explored a wide range of approaches to understanding social media discourse during the COVID-19 pandemic. This section reviews the most relevant studies that have contributed to these areas and highlights how the present work advances beyond existing methodologies. We have separately reviewed papers related to topic modeling, to sentiment analysis, to misinformation detection, and to bot identification, within the domain of this pandemic.

Table 1 presents the studies we reviewed regarding topic modeling.

Table 1: Reviewed studies associated with topic modeling in relation to COVID-19 tweets.

Year	Title of Paper	Authors	Models Used
2025	Temporal analysis of topic modeling output by machine learning techniques	Azizi, F., Vahdat-Nejad, H., Hajiabadi, H.	t-SNE, Fuzzy C-means, Topic Modeling
2025	Setting the misinformation agenda: Modeling COVID-19 narratives in Twitter communities	Unlu, A., Truong, S., Sawhney, N., Tammi, T.	BERT, BERTopic, Leiden algorithm

Year	Title of Paper	Authors	Models Used
2025	BI-SENT: bilingual aspect-based sentiment analysis of COVID-19 Tweets in Urdu language	Hashmi, E., Altaf, A., Anwar, M.W., Jamal, M.H., Bajwa, U.I.	LDA, NMF
2025	Analysis of Suicide-related Tweets During the COVID-19 Pandemic	Balasooriya, K.D.S., Rupasingha, R.A.H.M., Kumara, B.T.G.S.	LDA, PLSA
2025	Mitigating healthcare supply chain challenges under disaster conditions: a holistic AI-based analysis of social media data	Kumar, V.V., Sahoo, A., Balasubramanian, S.K., Gholston, S.	K-means clustering, Random Forest, Markov chain
2025	The dynamic response of coworking communities: unpacking resilience in times of disruption	Cochis, C., Bertolotti, F., Ungureanu, P.	LDA
2024	Identifying Crisis Response Communities in Online Social Networks for Compound Disasters: The Case of Hurricane Laura and COVID-19	Momin, K.A., Kays, H.M.I., Sadri, A.M.	BERTopic
2024	Machine Learning and Deep Learning Sentiment Analysis Models: Case Study on the SENT-COVID Corpus of Tweets in Mexican Spanish	Gomez-Adorno, H., Bel-Enguix, G., Sierra, G., Barajas, J.-C., Álvarez, W.	word2vec, doc2vec, BERT, Logistic Regression, Naive Bayes, SVM, MLP
2024	Sentiment analysis and topic modeling of COVID-19 tweets of India	Bhardwaj, M., Mishra, P., Badhani, S., Muttoo, S.K.	LDA, Decision Tree, Logistic Regression, Naive Bayes, SVM, AdaBoost, Random Forest
2024	Autoregressive Feature Extraction with Topic Modeling for Aspect-based Sentiment Analysis of Arabic as a Low-resource Language	Hashem Sweidan, A., El-Bendary, N., Elhariri, E.	BERT, XLNet, LDA, NMF, Bi-LSTM
2024	Deep learning-based user experience evaluation in distance learning	Sadigov, R., Yıldırım, E., Kocaçınar, B., Patlar Akbulut, F., Catal, C.	LSTM, word2vec, LDA
2024	Intelligent Recommendations Based on COVID-19 Related Twitter Sentiment Analysis and Fake Tweet Detection in Apache Spark Environment	Badawi, D.	Light GBM, Bi-GRU CapsNet, TD3, CVOA

Year	Title of Paper	Authors	Models Used
2024	Energy in the backseat? Investigating decarbonization dialogue in supply chain tweets during and after COVID-19	Shahzad, U., Sengupta, T., Rao, A., Sharma, G.D.	NTF (Nonnegative Tensor Factorization), LDA
2024	Detecting Reported Side Effects of COVID-19 Vaccines From Arabic Twitter (X) Data	Alhumayani, M.K., Alhazmi, H.N.	BTM (Biterm Topic Modeling), SVM
2023	Twitter discussions on breastfeeding during the COVID-19 pandemic	Jagarapu, J., Diaz, M.I., Lehmann, C.U., Medford, R.J.	VADER, NRC
2023	Evolution of COVID-19 tweets about Southeast Asian Countries: topic modelling and sentiment analyses	Mathayomchan, B., Taecharungroj, V., Wattanacharoensil, W.	LDA, VADER, NRC
2023	Exploring COVID-19 vaccine hesitancy and behavioral themes using social media big-data: a text mining approach	Yadav, H., Sagar, M.	K-means clustering, Random Forest
2023	Detecting and analyzing topics of massive COVID-19 related tweets for various countries	Azizi, F., Hajiabadi, H., Vahdat-Nejad, H., Khosravi, M.H.	LDA
2023	Topics in Antivax and Provax Discourse: Yearlong Synoptic Study of COVID-19 Vaccine Tweets	Zaidi, Z., Ye, M., Samon, F., Evans, J., Kashima, Y.	BERT
2023	Ensemble Deep Learning Framework for Situational Aspects-Based Annotation and Classification of International Student's Tweets during COVID-19	Hussain, S., Ayoub, M., Yu, Y., Moller, D.P.F., Weiyan, H.	Random Forest, LDA, PLSA
2022	Detection of Students' Problems in Distance Education Using Topic Modeling and Machine Learning	Alhazmi, H.	LDA
2022	#ChinaMustexplain: Global Tweets, COVID-19, and Anti-Black Racism in China	Ouassini, A., Amini, M., Ouassini, N.	LDA

Table 2 delineates the studies reviewed in the arena of sentiment analysis for COVID-19 tweets.

Table 2: Reviewed studies associated with sentiment analysis in relation to COVID-19 tweets.

Year	**Title of Paper**	**Authors**	**Models Used**
2025	Text Mining Data Summarization Approach for COVID-19 Using Hybrid Machine Learning Techniques	Lakshmi, C.S., Saxena, S., Kumar, B.S.	Bag-of-Words (BoW), TF-IDF, Word2Vec, Hybrid Machine Learning (HML)
2023	Sentiment Analysis on Twitters Big Data Against the Covid-19 Pandemic Using Machine Learning Algorithms	Sayed, A., Gomaa, M.M., Nazier, M.M.	Naive Bayes, Decision Tree, Logistic Regression, SVM
2023	ASAnalyzer: Attention based Sentiment analyzer for Real-world Sentiment Analysis	Hussain, K., Azhar, M., Lee, B., Affan, M., Ullah Khan, S.	CNN + Attention-Based BiGRU
2023	Fuzzy Based Text Quality Assessment for Sentiment Analysis	BenSassi, M., Abbes, M., Atigui, F.	Fuzzified classifier
2022	Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms	Bari, A., Heymann, M., Cohen, R.J., Dilorenzo, M., Coffee, M.	NLP sentiment analysis (specific models not detailed)
2022	Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic	Mahdikhani, M.	TF-IDF vectorizer, Bag of Words (BoW), Document embedding, Supervised ML algorithms
2022	Big Data Analysis for Healthcare Application using Minhash and Machine Learning in Apache Spark Framework	Albaldawi, W.S., Almuttairi, R.M., Manaa, M.E.	Minhash-LSH, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest
2021	Measuring the effectiveness of adaptive random forest for handling concept drift in big data streams	Alqabbany, A.O., Azmi, A.M.	Adaptive Random Forest (ARF), VADER
2020	A Data-Driven Method for Measuring the Negative Impact of Sentiment Towards China in the Context of COVID-19	Muñoz, L.M., Ramirez, M.F., Camargo, J.E.	VADER
2020	Mining of Social Media on Covid-19 Big Data Infodemic in Indonesia	Binsar, F., Mauritsius, T.	Support Vector Machine, Random Forest, Naïve Bayes

Year	Title of Paper	Authors	Models Used
2020	Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System	Zhang, X., Saleh, H., Younis, E.M.G., Sahal, R., Ali, A.A.	Decision Tree, Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, TF-IDF, n-gram

Table 3 describes the studies reviewed in the sphere of outlier detection in the context of COVID-19 tweets.

Table 3: Reviewed studies associated with anomaly detection in relation to COVID-19 tweets.

Year	Title of Paper	Authors	Models Used
2023	Automatic identification and explanation of root causes on COVID-19 index anomalies	Sufi, F.K.	CNN
2022	An Anomaly Detection Framework for Twitter Data	Kumar, S., Khan, M.B., Hasanat, M.H.A., AlTameem, A., AlKhathami, M.	NMF, LDA, Sentence Transformer, K-means

Table 4 explains the studies about bot identification with regards to COVID-19 tweets.

Table 4: Reviewed studies associated with bot identification in relation to COVID-19 tweets.

Year	Title	Authors	Models Used
2025	Tracing the dynamics of misinformation and vaccine stance in Finland amid COVID-19	Unlu, A., Truong, S., Sawhney, N., Sivelä, J., Tammi, T.	FinBERT, Botometer, BERTopic
2024	The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker	Dadkhah, S., Zhang, X., Weismann, A.G., Firouzi, A., Ghorbani, A.A.	BERT-based models, Classical ML algorithms, DBSCAN, YAKE
2024	Long-term assessment of social amplification of risk during COVID-19: challenges to public health agencies amid misinformation and vaccine stance	Unlu, A., Truong, S., Sawhney, N., Sivelä, J., Tammi, T.	BERT models, Botometer, Social Network Analysis
2023	A Comparative Study of Bot Detection Techniques With an Application in Twitter Covid-19 Discourse	Antenore, M., Camacho Rodriguez, J.M., Panizzi, E.	Botometer v3, Digital Fingerprint-based models, Social Fingerprint-based methods

Year	Title	Authors	Models Used
2022	Online Conspiracy Groups: Micro-Bloggers, Bots, and Coronavirus Conspiracy Talk on Twitter	Greve, H.R., Rao, H., Vicinanza, P., Zhou, E.Y.	Biterm Topic Model (BTM), Event-history analysis
2022	Using Natural Language Processing to Explore "Dry January" Posts on Twitter: Longitudinal Infodemiology Study	Russell, A.M., Valdez, D., Chiang, S.C., Lin, H.-C., Massey, P.M.	TF-IDF, K-means clustering, PCA, Latent Dirichlet Allocation (LDA), VADER sentiment analysis, Botometer
2021	Hunting Conspiracy Theories During the COVID-19 Pandemic	Moffitt, J.D., King, C., Carley, K.M.	BERT (Bidirectional Encoder Representations from Transformers)
2021	Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter	Qureshi, K.A., Malick, R.A.S., Sabih, M., Cherifi, H.	CATBoost, RNN, ML and DL models, Complex Network Analysis

An analysis of the four tables reveals evolving trends in the computational study of COVID-19-related Twitter discourse from 2020 to 2025. The reviewed studies collectively trace the maturation of social media analytics methodologies during and after the pandemic.

From 2020 through 2022, topic modeling studies relied heavily on Latent Dirichlet Allocation (LDA), which remained the dominant approach for discovering latent themes in COVID-19 tweets. During this period, LDA appeared in nearly every study (e.g., Ouassini et al., 2022; Alhazmi, 2022), often coupled with clustering techniques like K-means or PLSA. However, from 2023 onward, a marked methodological shift occurred. Transformer-based and hybrid models, such as BERTopic, BERT, and Nonnegative Tensor Factorization (NTF), gained prominence (e.g., Momin et al., 2024; Shahzad et al., 2024; Unlu et al., 2025). This transition mirrors a broader trend in natural language processing: a move from purely probabilistic topic inference toward contextual embedding-based semantic modeling, which captures nuanced meanings and co-occurrence patterns more effectively. Furthermore, post-2023 studies show a clear emphasis on domain-specific adaptation, such as multilingual and aspect-based topic extraction (e.g., Hashmi et al., 2025), reflecting global engagement with the pandemic narrative.

Sentiment analysis in the COVID-19 context began in 2020 with the application of traditional machine learning techniques (Naive Bayes, SVM, Decision Tree, Random Forest) and rule-based lexicon models like VADER (Muñoz et al., 2020). By 2022, researchers had begun integrating hybrid architectures, feature embeddings (TF-IDF, Word2Vec), and deep learning frameworks (BiGRU, CNN), signaling a shift toward more context-sensitive emotional classification. The 2023–2025 period saw an explosion in hybrid and attention-based networks (e.g., Hussain et al., 2023; Lakshmi et al., 2025), combining linguistic, syntactic, and semantic cues. This evolution reflects a movement from coarse-grained sentiment polarity detection to fine-grained emotion recognition and aspect-based sentiment categorization, particularly for multilingual and domain-specific tweet corpora. Additionally, sentiment analysis increasingly became a precursor or integrated component in misinformation detection and topic modeling workflows.

Compared with topic and sentiment analysis, outlier or anomaly detection appears less frequently in the literature, indicating that it is an emergent but not yet mature domain. The earliest appearance in this dataset is in 2022 (Kumar et al.), where K-means and Sentence Transformers were applied to detect unusual tweet behavior patterns. By 2023, CNN-based frameworks (Sufi, 2023) began being employed to automatically explain the causes of anomalies, indicating a growing sophistication in identifying contextual outliers linked to misinformation, spam, or social shocks. Despite limited representation, this domain reflects an important methodological pivot: researchers are starting to view anomaly detection not merely as noise elimination but as a means to reveal strategic disinformation and atypical discourse dynamics.

Bot detection research displays a strong upward trajectory from 2021 to 2025. Early studies (e.g., Moffitt et al., 2021; Qureshi et al., 2021) combined network analysis with classical machine learning (RNN, CATBoost) to classify bot-generated tweets. Starting in 2022, integration of Botometer and transformer models (BERT, FinBERT) became standard practice (Russell et al., 2022; Unlu et al., 2024, Unlu et al., 2025). The later studies increasingly fuse content analysis with graph-based features, suggesting a convergence between natural language processing and social network analytics. By 2025, bot identification had evolved into a hybrid intelligence task, analyzing both textual semantics and diffusion patterns, underscoring its critical role in mitigating coordinated misinformation campaigns. The use of BERTopic and FinBERT in misinformation-bot joint analyses (Unlu et al., 2025) reflects the rising sophistication and integrative nature of this research stream.

Across all four domains, 2020–2021 marks the exploratory phase dominated by statistical and classical machine learning models. The 2022–2023 phase signifies methodological diversification, introducing hybrid pipelines, multilingual corpora, and multi-task learning. From 2024 onward, a consolidation occurs around transformer-based architectures (BERT, BERTopic, FinBERT) and explainable AI techniques, reflecting both computational advancement and increasing concern with interpretability, scalability, and misinformation mitigation. Importantly, many 2025 studies adopt multi-domain frameworks, for instance, linking topic modeling with misinformation detection or sentiment analysis with bot identification, indicating a holistic approach to understanding the pandemic's information ecosystem.

2.1 How different ideas compare from others

Across the reviewed literature, each paper contributes a distinct idea or methodological perspective shaped by its objectives, data, and analytical scope. Early works (2020–2021) emphasize foundational analyses using statistical models such as LDA, Naive Bayes, and SVM to uncover general topics or sentiments about the pandemic. These studies often treat Twitter data as a large, unstructured text source for descriptive analytics. From 2022 onward, researchers introduce more specialized or hybrid frameworks, combining topic modeling with feature embeddings, or sentiment analysis with temporal or aspect-based components—to capture evolving discourse patterns and nuanced emotional expressions. By 2023–2024, ideas become more interdisciplinary: some studies integrate social network analysis to identify misinformation and conspiracy clusters, while others employ multilingual or domain-specific models like BERT, BERTopic, and FinBERT to enhance contextual accuracy. In 2025, the most innovative works shift toward explainability and integration, linking topic evolution with misinformation agendas, combining bot detection with transformer-driven narrative analysis, or applying hybrid AI to trace sentiment shifts in healthcare and public communication. Thus, each study differs by its conceptual angle, linguistic, behavioral, or structural. They progress from isolated single-task models to complex, context-aware systems that jointly examine semantics, emotion, and authenticity within pandemic-related social media discourse.

2.2 How our ideas compare with the ideas in reviewed papers

Compared to prior studies, our proposed framework endeavors a methodological and conceptual advancement in the analysis of COVID-19 Twitter discourse. Our approach transcends fragmentation and unites four major analytical axes, latent topic discovery, affective quantification, anomaly detection, and bot identification. In contrast to many prior studies that rely on manually annotated and labeled datasets, our design attempts to be an unsupervised and zero-shot approach which posits itself to accomplish interpretability and scalability without human intervention. Methodologically, we progress beyond traditional keyword- or frequency-based clustering by applying HDBSCAN directly on high-dimensional transformer embeddings. While existing sentiment studies often use single-layer polarity models, our dual-layer affective estimation adds a second dimension of emotion classification and enables a deeper psychological understanding of pandemic discourse.

The integration of Isolation Forests for anomaly detection and Botometer for automated account verification introduces a quality-control layer absent in earlier works. It empowers analytical insights to be derived from genuine human communication rather than synthetic amplification. While prior research tends to examine what people said or how they felt, our pipeline establishes a holistic, unsupervised ecosystem that reveals what was said, how it was felt, whether it was genuine, and when it changed, positioning our contribution as a comprehensive evolution of COVID-19 social media analytics.

3 Methodology

This section is organized to provide a clear, logical, and reproducible roadmap of the study's approach. It begins by outlining the research objectives to clearly articulate the goals and questions guiding the work. This is followed by the dataset description which establishes the crucial foundation of the study by detailing the source, attributes, and collection process of the data used for the analysis. The section culminates in the proposed approach where the specific sequence of methods is detailed and justified.

3.1 Research Objectives

Our method aims to enhance the efficiency and effectiveness of COVID-19 public health monitoring by automatically identifying, classifying, and verifying the key topics, sentiments, and sources present in real-time social media discourse. To achieve this goal, our method has set the following research objectives (ROs) and prepares answers to the research questions (RQs):

RQ1: How can unsupervised methods reveal the latent structure and subtopics within the high-volume, diverse corpus of COVID-19 tweets?

RO1: We intend to perform granular topic modeling. The objective is to move beyond simple keyword counting to identify specific, nuanced subtopics being discussed in the COVID-19 conversation, providing a structural map of the public discourse. We apply HDBSCAN clustering directly to the high-dimensional transformer-generated tweet embeddings to isolate dense, semantically coherent clusters. Subsequently, the BERTopic framework is employed to process these clusters, automatically generating interpretable topic keywords and extracting representative tweets to summarize each latent subtopic. The primary outcome is a set of well-defined, easily interpretable subtopics, demonstrating the diversity and underlying structure of the dataset. If timestamps are available, a secondary outcome is the identification of temporal topic evolution and event bursts, revealing how the public conversation changed over time.

RQ2: Can pre-trained, zero-shot models accurately and automatically quantify the emotional tone and polarity of the public's response to COVID-19 events?

RO2: We endeavor to automate dual-layer affective estimation. The objective assigns a dual layer of affective labels (sentiment polarity and specific emotion) to every tweet, which is critical for understanding the public's psychological state during the pandemic. This process is designed to operate without any manual labeling (zero-shot/transfer learning). The specialized `cardiffnlp/twitter-roberta-base-sentiment` model is used to classify each tweet's polarity (positive, negative, and neutral). The `j-hartmann/emotion-english-distilroberta-base` model is utilized to tag tweets with a wider range of specific emotions (e.g., joy, anger, sadness). The resulting tweets are enriched with automated sentiment and emotion scores. The main contribution is the aggregation of sentiment trends over time and across the topics identified in RO1 and provide a comprehension of the emotional geography of the pandemic response.

RQ3: What unsupervised methods can effectively identify statistically anomalous tweets that may represent misinformation, spam, or rare but significant events?

RO3: Our aim is to implement unsupervised detection of anomalous content. We intend to detect statistically distant data points that deviate significantly from the semantic patterns of the main tweet corpus to identify potential content integrity issues and rare but important signals. The Isolation Forest algorithm is applied to the high-dimensional tweet embeddings. This algorithm isolates data points that require few splits to separate them from the rest of the dataset which indicates their anomalous nature. This step generates an outlier score for every tweet. The outcome is the identification and flagging of anomalous tweets (potential spam, misinformation, and highly unique and irregular content) for further investigation to increase the dataset's validity for subsequent analyses.

RQ4: How can pre-trained models be leveraged to identify and quantify the presence of automated (synthetic) accounts contributing to the COVID-19 discussion?

RO4: We aim to conduct tweet-level bot identification. The objective here focuses on assessing data integrity and quantifies the extent to which automated accounts (bots) are participating in the conversation. This distinguishes genuine human content from synthetic amplification. The widely-used pre-trained Botometer model is applied to tweets in the dataset to analyze numerous features for each tweet. The primary contribution is a Bot Score metric associated with each tweet's author. This allows for the quantification and removal (or statistical weighting) of automated content, ensuring that the findings from the topic and sentiment analyses (RO1 and RO2) accurately reflect human public opinion.

3.2 Dataset Description

This study utilizes a large-scale dataset known as COVID-19 Twitter Dataset. A representative tweet from the dataset is given in Figure 1. It is a dataset of English-language tweets collected globally during the COVID-19 pandemic. The dataset was constructed in three distinct temporal phases, following the World Health Organization's declaration of COVID-19 as a global pandemic. Tweets were streamed directly from Twitter's public API at an average rate of nearly ten thousand per day, ensuring broad temporal and geographical coverage. The collection strategy was designed to capture evolving public discourse, behavioral trends, and informational flow related to COVID-19 across different phases of the pandemic.

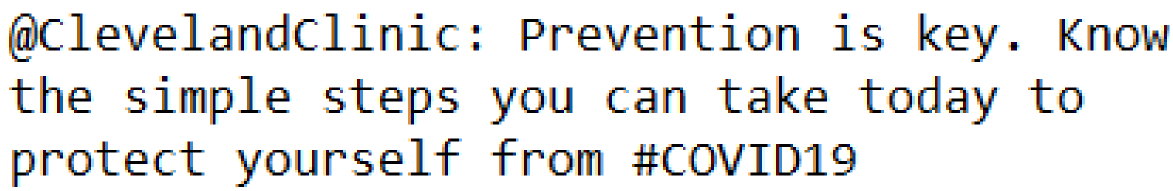

Figure 1: A Tweet Instance from the Dataset

The first phase dataset was composed of approximately 235,240 tweets collected between April 19th and June 20th, 2020, representing the early global response to the pandemic. After a brief hiatus, data collection resumed for the second phase, which included about 320,316 tweets gathered between August 20th and October 20th, 2020, corresponding to the period when the pandemic reached heightened fatality levels. The third phase dataset, comprising nearly 489,269 tweets, was collected between April 26th and June 27th, 2021, reflecting renewed public engagement during the emergence of vaccine deployment and subsequent infection waves. Altogether, the compiled corpus encompasses more than one million tweets distributed across three key stages of the global pandemic timeline.

Tweets were collected using a targeted set of COVID-19-related hashtags, including #covid19, #coronavirus, #covid, #covaccine, #lockdown, #homequarantine, #quarantinecenter, #socialdistancing, #stayhome, and #staysafe. A hashtag-based approach was conducive for the relevance of the tweets to pandemic-related discussions. The dataset includes a variety of tweet attributes such as tweet ID, timestamp, source application, original text, language, favorite count, retweet count, author handle, hashtags, and user mentions. Location is also mentioned wherever extricable.

Comprehensive data pre-processing was performed using a custom-built function based on the Natural Language Toolkit (NLTK). The preprocessing pipeline standardized all text to lowercase and removed extraneous elements including white spaces, numbers, special symbols, ASCII characters, URLs, punctuation, and stopwords. To ensure lexical consistency, all instances of the word "covid" were replaced with "covid19." Stemming was applied to reduce words to their root form, enhancing the quality of textual analysis. This procedure produced a clean and normalized textual corpus suitable for advanced natural language processing tasks and large-scale computational studies.

3.3 Proposed Approach

The aim is to develop and implement a novel pipeline for real-time COVID-19 public health monitoring through social media discourse analysis. The goal is to utilize and integrate unsupervised machine learning, pre-trained transformer models, and anomaly detection algorithms to identify latent topics, quantify affective responses, and assess content integrity. The research objectives that have been delineated are addressed through complementary computational techniques that are operated on high-dimensional semantic representations of Twitter data. The summarized and condensed process is shown in Figure 2.

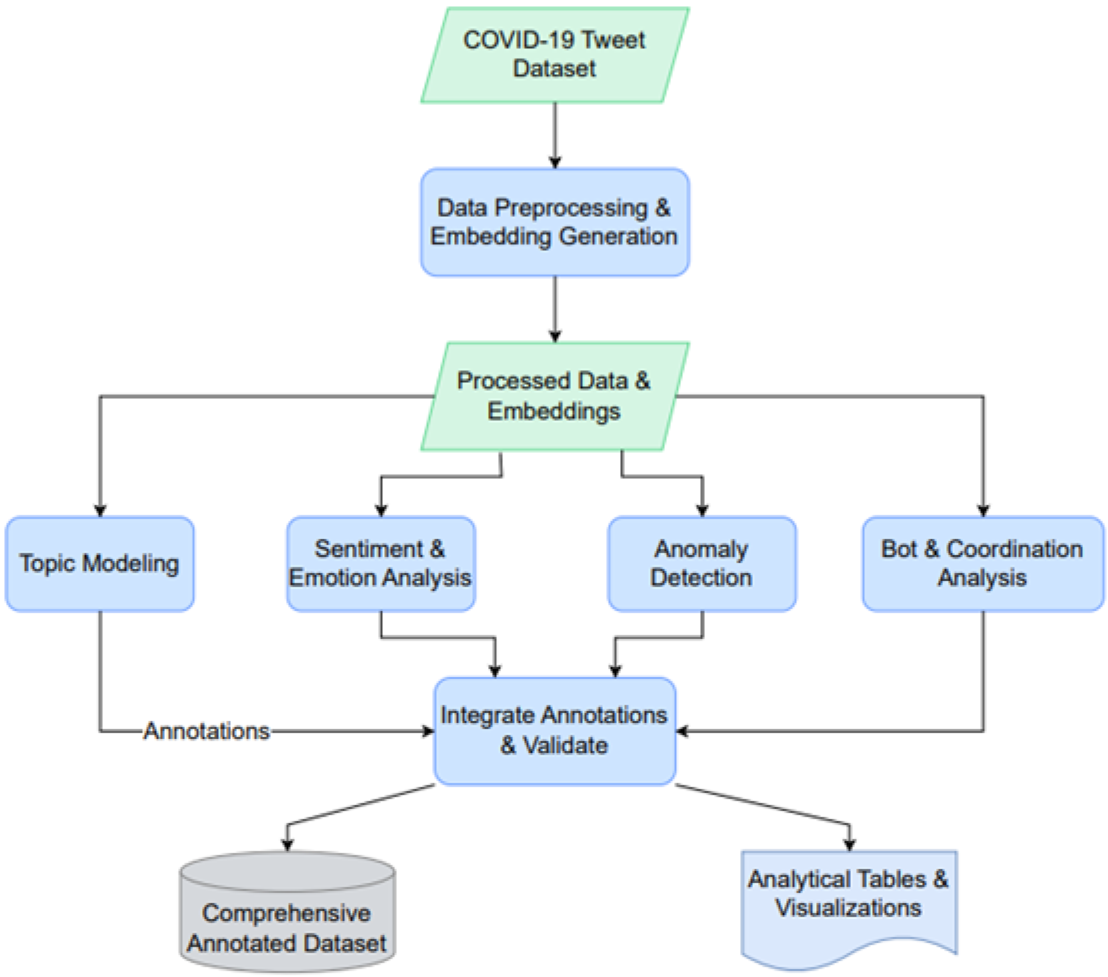

Figure 2: Overarching Diagram of the Proposed Approach

Data Preprocessing

The above-communicated dataset of COVID-19-related tweets is preprocessed before further tasks can be carried out. A comprehensive preprocessing pipeline is implemented wherein exact and near-duplicate tweets are removed using hash-based matching to guarantee data deduplication. Although the dataset mentions that tweet text normalization has been carried out, it is reinforced through conversion to lowercase, URL removal, and handling of special characters. Hashtags and mentions are preserved where semantically relevant. Language filtering is applied to retain English-language tweets only, where verification is performed using the `langdetect` library. Metadata extraction is executed to preserve timestamps, user identifiers, and engagement metrics including retweets and likes for subsequent temporal and network analyses. Stop words are intentionally retained during initial embedding generation to preserve contextual information for transformer models, as modern contextualized embeddings are designed to utilize these tokens for semantic understanding.

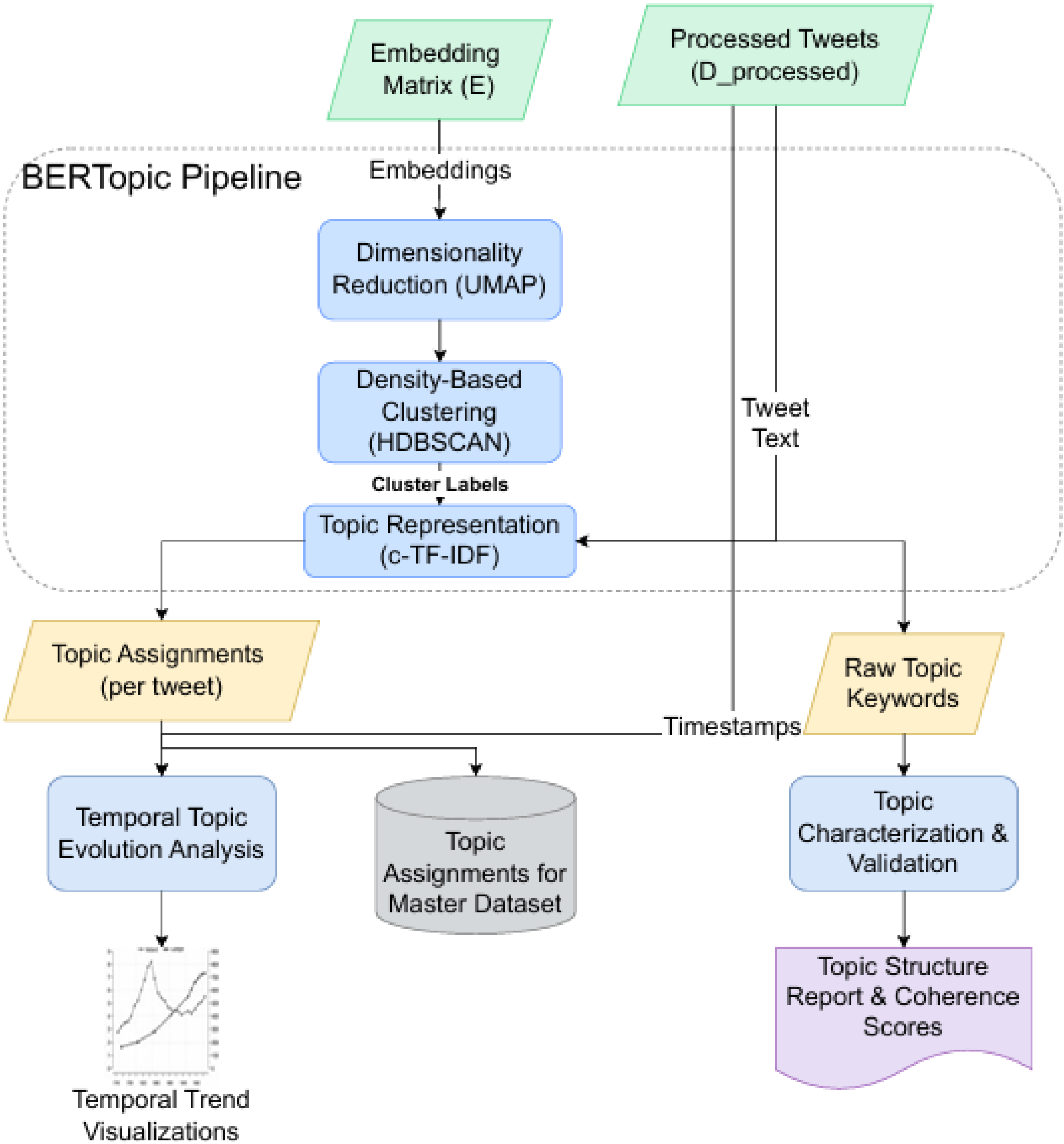

Figure 3: Proposed Approach for Topic Modeling

Feature Extraction

All tweets are encoded into high-dimensional semantic representations using `sentence-transformers/all-MiniLM-L6-v2`. It is a pre-trained transformer model which is optimized for semantic similarity tasks. This model is configured to generate 384-dimensional dense vector embeddings that capture contextual meaning beyond keyword matching. The model is selected to balance computational efficiency with semantic fidelity, making it suitable for large-scale real-time processing. For each tweet t, an embedding vector is obtained. This is the fundamental feature representation for all subsequent analyses including topic modeling, anomaly detection, and semantic clustering.

Topic Discovery and Temporal Evolution

Here, unsupervised methods are utilized to reveal the latent structure and subtopics within the high-volume, diverse corpus of COVID-19 tweets. The overview of this process is shown in Figure 3.

Density-Based Clustering HDBSCAN is applied directly to the tweet embedding space to identify semantically coherent clusters without requiring pre-specified cluster counts. The algorithm is configured with a minimum cluster size of 50 tweets per cluster to ensure statistical robustness, while a `min_samples` parameter of 15 is set to control cluster conservativeness. Euclidean distance is employed as the metric in embedding space, and the Excess of Mass method (EOM) is selected for cluster selection. Through this configuration, dense regions in the high-dimensional space are identified by HDBSCAN, while low-density points are classified as noise, effectively filtering semantically isolated or incoherent tweets.

Topic Modeling with BERTopic The BERTopic framework is employed to transform HDBSCAN clusters into interpretable topics. Within the BERTopic pipeline, dimensionality reduction is performed using UMAP (Uniform Manifold Approximation and Projection), where embeddings are reduced to 5 dimensions for visualization while local and global structure are preserved. Document representation is achieved through calculation of cluster centroids in embedding space, where tweets within each cluster are represented by their collective center. Topic representation is generated using class-based TF-IDF (c-TF-IDF), through which the most distinctive keywords for each cluster are extracted to generate topic labels. Representative documents are identified for each topic by BERTopic, where exemplar tweets closest to the cluster centroid are selected.

Through this process, a set of k interpretable topics is produced, where each topic is characterized by top-n keywords ($n = 10$), 3–5 (average 4) representative tweets, and topic coherence scores that are calculated to assess semantic consistency. A structural map of COVID-19 discourse subtopics and their semantic relationships is generated as the primary output of this research objective.

Dual-Layer Affective Estimation

Here, pre-trained zero-shot models are used to accurately quantify the emotional tone and polarity of the public's response to COVID-19 events. A two-tier affective labeling system is implemented using domain-specialized pre-trained transformer models, where we aim to do away with task-specific fine-tuning. The approach is graphically illustrated in Figure 4.

Sentiment Polarity Classification The `cardiffnlp/twitter-roberta-base-sentiment` model is employed for sentiment analysis. This is a RoBERTa-based architecture that has been previously fine-tuned on approximately 58 million tweets. Three-class sentiment classification is performed to categorize tweets into Positive, Neutral, or Negative categories. For each tweet t, a probability distribution $P(s \mid t)$ over sentiment classes is output by the model. The sentiment label is assigned as the argument that maximizes $P(s \mid t)$, while the confidence score is recorded as the maximum value of $P(s \mid t)$.

Granular Emotion Classification The `j-hartmann/emotion-english-distilroberta-base` model is utilized for granular emotion detection. This is a DistilRoBERTa-based architecture which is fine-tuned on six emotion datasets. Multi-class emotion classification is performed across seven categories including anger, disgust, fear, joy, neutral, sadness, and surprise. For each tweet, a probability distribution $P(e \mid t)$ is obtained, from which the dominant emotion label and associated confidence score are assigned.

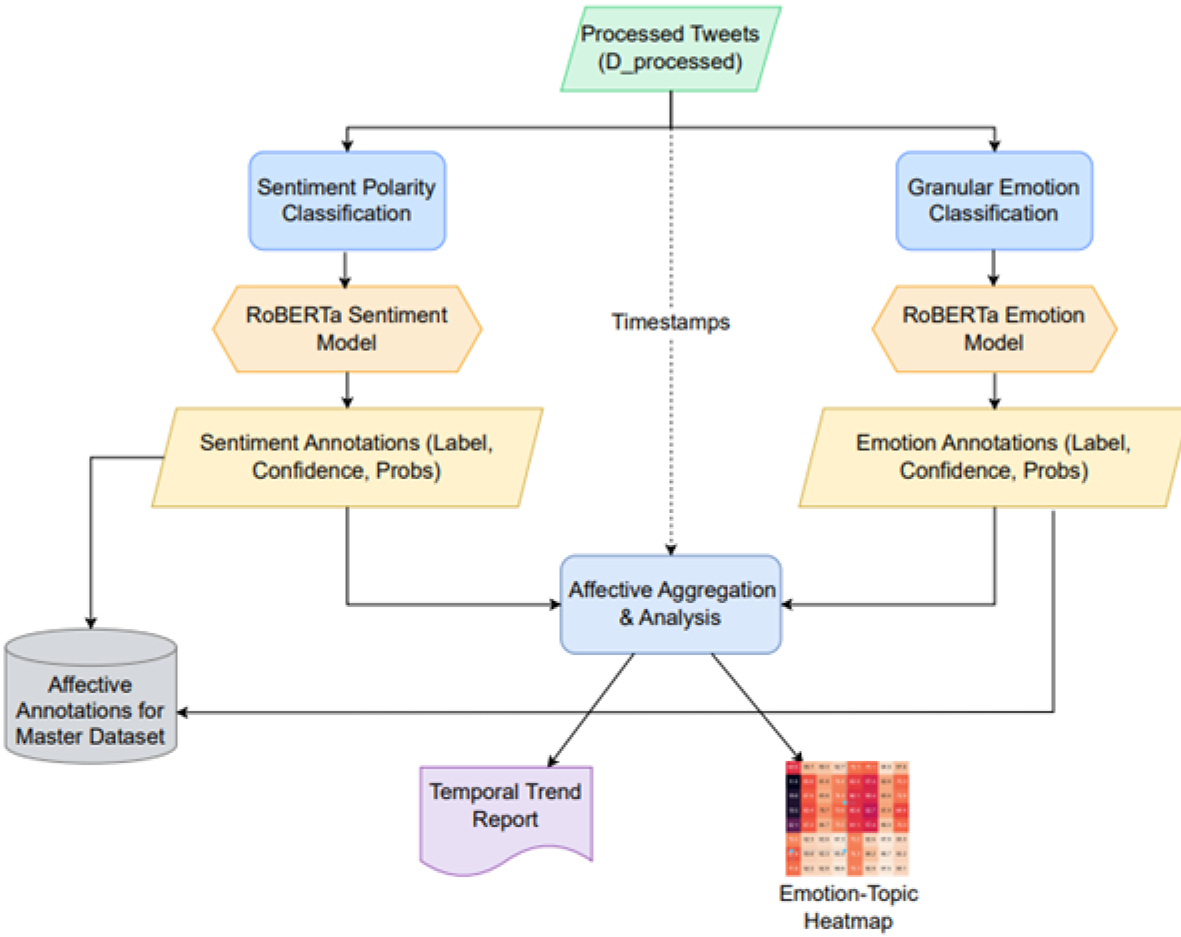

Figure 4: Illustration of the Sentiment Analysis Module

Affective Aggregation and Analysis Multiple levels of affective analysis are computed to understand the emotional landscape of COVID-19 discourse. Topic-level sentiment profiles are generated through calculation of average sentiment distribution across tweets in each topic. Temporal sentiment trends are tracked by computing daily and weekly average sentiment polarity scores across the dataset timeline. Emotion-topic heatmaps are constructed through cross-tabulation of emotion distributions across all identified topics, revealing the affective character of different discussion themes. Event correlation analysis is performed through alignment of sentiment shifts with documented real-world COVID-19 events such as policy announcements and case surges. This empowers a comprehensive emotional mapping of pandemic discourse without any manual annotation effort.

Unsupervised Detection of Anomalous Content

Here, unsupervised methods are utilized to identify statistically anomalous tweets that may represent misinformation, spam, and rare but significant events. Figure 5 depicts this module diagrammatically.

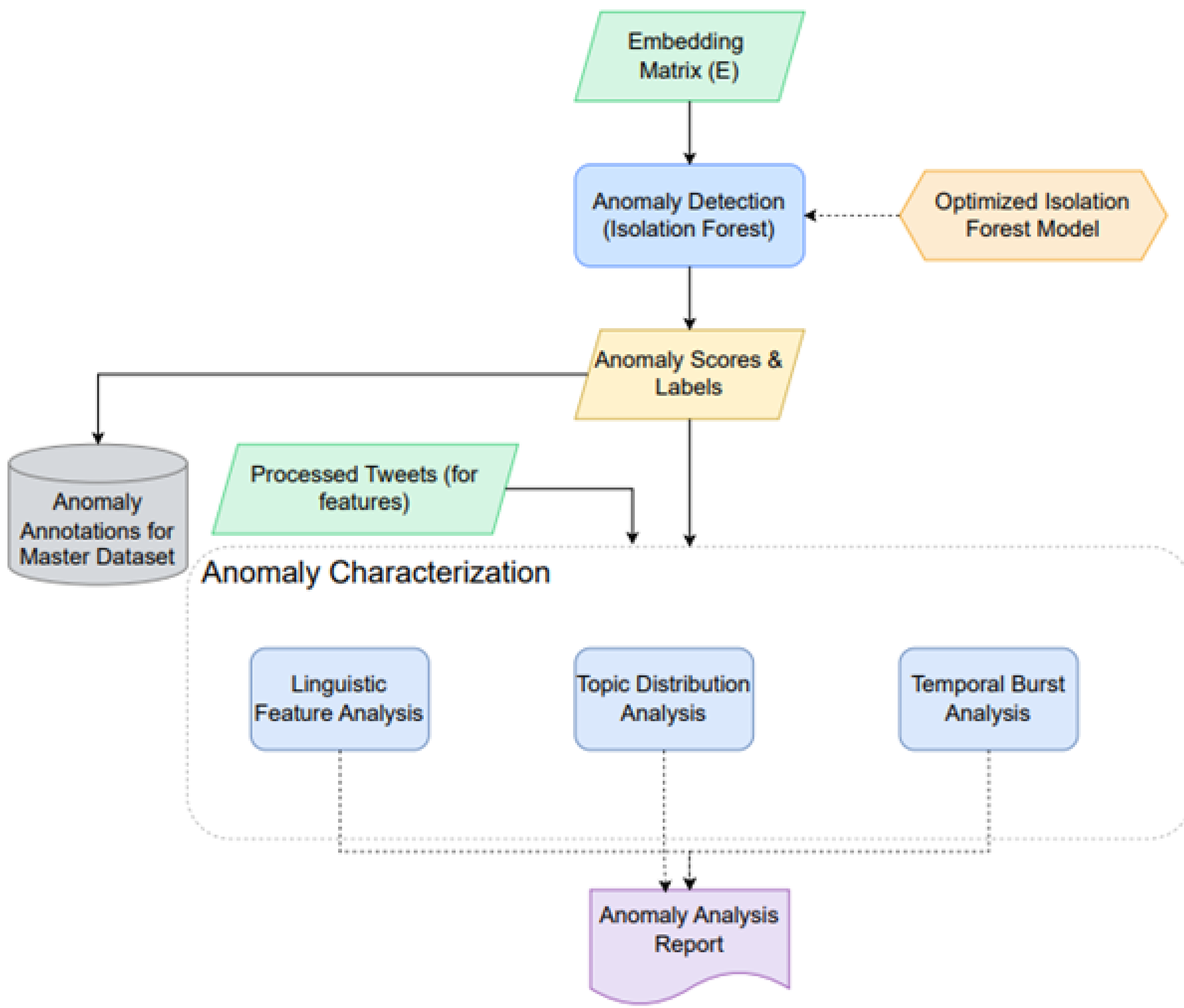

Figure 5: Diagrammatic Representation of our Outlier Detection Process

Isolation Forest for Outlier Detection Isolation Forest is an ensemble-based anomaly detection algorithm. This is applied directly to tweet embeddings to identify statistically deviant content. The algorithm is operated on the principle that anomalies are few and different. They require fewer random partitions to isolate from the main data distribution. The implementation is configured with a certain number of isolation trees.

The number of isolation trees is determined based on the stability of anomaly scores. Initially, a moderate number of trees, such as 100, is configured, and the Isolation Forest is applied to the dataset to generate anomaly scores for all points. The stability of these scores is then assessed by computing the variance of scores across multiple runs or subsets of trees and by measuring the consistency of point rankings between successive runs. The number of trees is incrementally increased. The process is repeated until the variance or ranking changes fall below a predefined threshold so that additional trees do not significantly alter the anomaly scoring performance.

Contamination is set to `auto` to allow automatic threshold determination based on data distribution. A maximum sample size of 256 is specified for each tree, and a random state of 42 is set to ensure reproducibility. For each tweet t, an anomaly score $a_t \in [-1, 1]$ is produced by the algorithm, where negative values are designated to indicate anomalies. A binary label is assigned to classify each tweet as either an outlier (1) or inlier (0) based on the computed anomaly score.

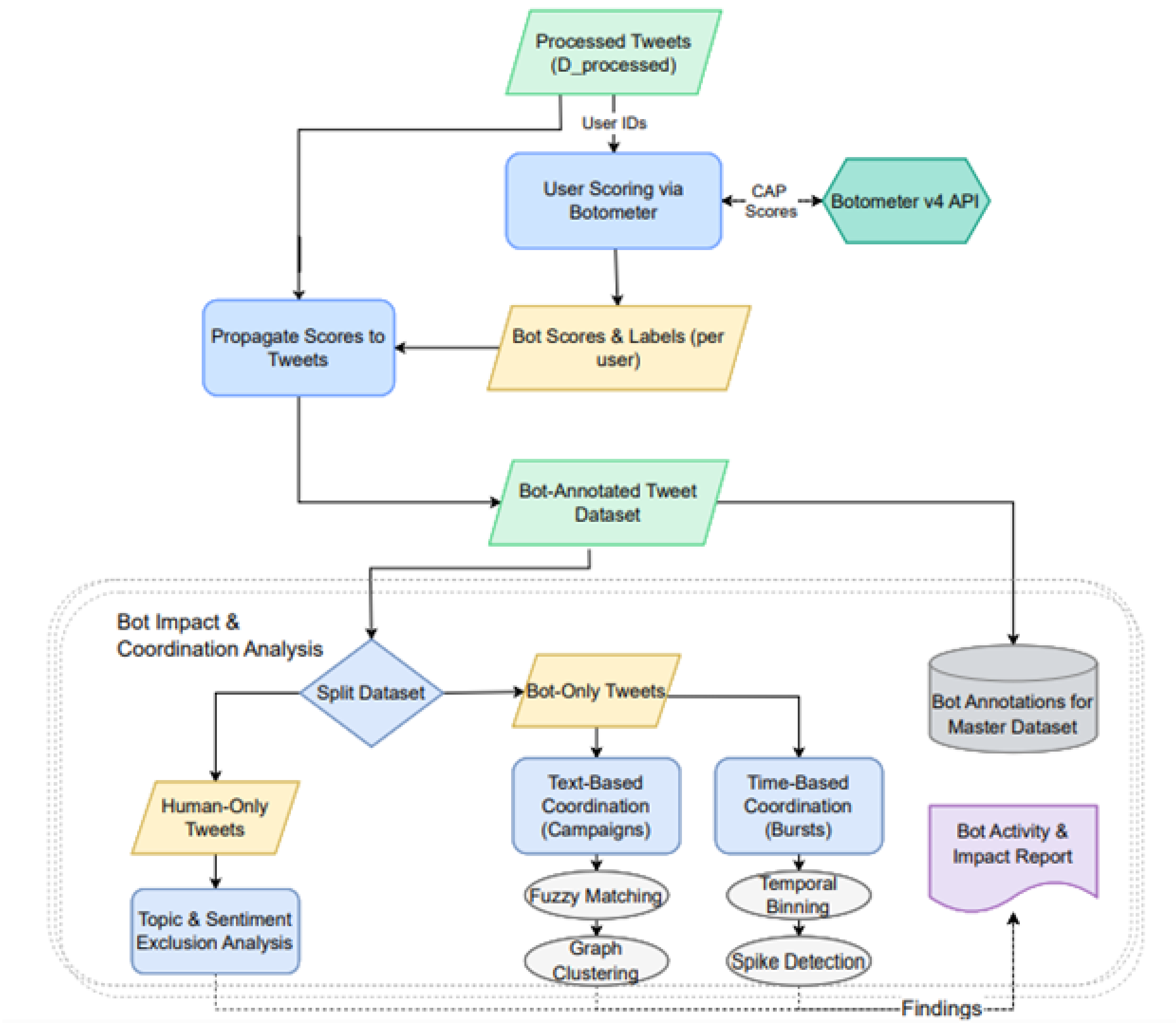

Figure 6: Proposed Bot Detection Paradigm

Botometer-Based Bot Score Assignment

Anomaly Interpretation and Validation Detected anomalies are characterized through a validation approach. Topic distribution analysis is performed to compare topic membership between anomalous and normal tweets to find out whether anomalies cluster in specific discourse areas. Linguistic feature analysis is executed through examination of text length, special character frequency, and URL density to identify structural characteristics of anomalous content. Temporal clustering analysis is implemented to identify anomaly bursts that coincide with known misinformation campaigns or unusual events. An anomaly-flagged dataset is generated as output, enabling filtered analysis and integrity assessment for subsequent research applications.

Bot Detection

Here, pre-trained models are leveraged to identify and quantify the presence of automated accounts contributing to the COVID-19 discussion. Figure 15 represents our bot detection paradigm diagrammatically.

Botometer (v4) is a machine learning framework that is developed by Indiana University. It is employed to analyze multiple account-level and behavioral features to estimate bot likelihood. Tweet content and linguistic features is analyzed for each tweet text. For each unique tweet t in the dataset, a complete automation probability score ranging from 0 to 1 is returned by Botometer. It shows the likelihood that the tweet is automated. A conservative threshold of CAP ≥ 0.75 is applied to classify tweets as likely generated by bots. This threshold is selected following established literature recommendations for balanced precision-recall performance in social media studies.

Dataset Weighting Strategy To ensure that findings reflect genuine human discourse rather than automated amplification, multiple analytical strategies are implemented. Exclusion analysis is performed where topic and sentiment analyses are conducted both with and without bot-generated tweets for comparative assessment. Weighted aggregation is applied where human-generated tweets are weighted by $(1 - \text{CAP})$ in sentiment trend calculations to proportionally diminish bot influence. Bot activity characterization is conducted through analysis of topic and sentiment distributions within bot-generated content to identify coordinated amplification patterns and inauthentic coordination. A bot-score annotated dataset is produced as output, enabling authentic public opinion analysis by distinguishing human-generated discourse from automated content.

The findings are guided to reflect genuine human discourse rather than automated amplification. Multiple analytical strategies are implemented for this. Exclusion analysis is performed by conducting topic and sentiment analyses both with and without bot-generated tweets to assess differences. For example, in one dataset, human tweets primarily discussed vaccine availability and hospital capacity, whereas including bots introduced a cluster promoting a conspiracy narrative which may showcase bot-driven content. Weighted aggregation is also applied where human-generated tweets are weighted by $(1 - \text{CAP})$ in sentiment trend calculations to proportionally diminish bot influence. For example, a human tweet with a CAP of 0.2 contributes 0.8 to the aggregated score, whereas a likely bot tweet (say CAP $= 0.9$) contributes only minimally (weight $= 0.1$) if mistakenly included. Bot activity characterization is conducted through analysis of topic and sentiment distributions within bot-generated content to identify coordinated amplification patterns. For example, repeated messaging across multiple accounts using hashtags like #VaccineDanger can reveal inauthentic coordination that is not present in human discourse. Using text embeddings or fuzzy string matching, tweets that are nearly identical and share repeated phrases and hashtags can be flagged.

Integration and Statistical Analysis

All analytical components are integrated into a unified pipeline using Python 3.9+ with multiple specialized libraries. `scikit-learn` is employed for clustering and anomaly detection, `sentence-transformers` is utilized for embedding generation, BERTopic is applied for topic modeling, the `transformers` library is used for sentiment and emotion models, and the Botometer API is accessed for bot detection.

Statistical significance of temporal trends is assessed using Mann-Kendall tests for monotonic trends, while changepoint detection algorithms are applied to identify abrupt shifts in discourse patterns. Topic coherence is measured using both automated metrics, specifically C_v coherence scores, and human evaluation of topic interpretability is conducted on a random sample of 20 topics to validate assessments.

All visualizations and temporal analyses are conducted using `matplotlib`, `seaborn`, and `plotly` libraries, where interactive exploration of high-dimensional results is enabled through dynamic plotting capabilities. The algorithm for the process is shown below.

Algorithm 1 COVID-19 Tweet Analysis Pipeline

Input: Raw COVID-19 tweet dataset ($tweet\_id, text, timestamp, user\_id, retweet\_count, favorite\_count, hashtags, mentions$)

Output: Annotated dataset with topics, sentiments, emotions, anomaly scores, bot scores, and analytical reports

1: **Data Preprocessing:** Load and merge tweets, remove duplicates, normalize text, filter English tweets, extract metadata.
2: **Semantic Embedding:** Generate embeddings using sentence-transformers/all-MiniLM-L6-v2.
3: **Topic Discovery:** Apply HDBSCAN on embeddings to obtain clusters and noise points.
4: **Topic Modeling:** Use BERTopic with UMAP for dimensionality reduction; extract top keywords and representative tweets.
5: **Temporal Topic Analysis:** Aggregate topics into time bins, compute prevalence, and detect changepoints.
6: **Sentiment and Emotion Classification:** Assign sentiment and emotion labels using transformer-based classifiers; compute topic-level distributions and temporal trends.
7: **Anomaly Detection:** Fit Isolation Forest on embeddings; compute anomaly scores and identify outliers.
8: **Bot Detection and Impact Analysis:** Estimate bot likelihood; adjust sentiment and topic statistics by bot influence; detect coordinated campaigns.
9: **Validation:** Apply statistical tests (Mann–Kendall, PELT) and validate topic coherence against human ratings.

4 Results

The results of our study show structural patterns in activity, thematic organization, and emotional characterization within the dataset. The analyses provide an integrated view of temporal behavior, topic prominence, affective distribution, and relational structure across the conversation space. Each component highlights distinctive features that shape the dynamics of the discourse.

We performed the implementation in two runs. The first run provided some results which necessitated improvement, and an improved second run was performed and its results were analyzed. Here the need for further improvements was encountered and these improvements are planned for a future iteration.

4.1 Results for the first run

Figures 7 through ten describe the performance of the first run.

According to Figure 7, the temporal activity analysis reveals distinct behavioral signatures across the observation period, which spans April 22 to June 22, 2020. The line charts demonstrate a pronounced escalation in tweet volume during the first half of May, marking this interval as a phase of intensified public engagement. Despite the sharp rise in activity, the sentiment trajectory remains stable, displaying only a marginally positive drift. This steadiness suggests that the increase in discourse was not accompanied by substantial emotional volatility. A single extreme anomaly emerges on May 15, indicating a sudden irregularity in posting behavior. Bot activity, however, stays consistently negligible, implying that the conversation dynamics were driven primarily by organic user participation rather than automated amplification.

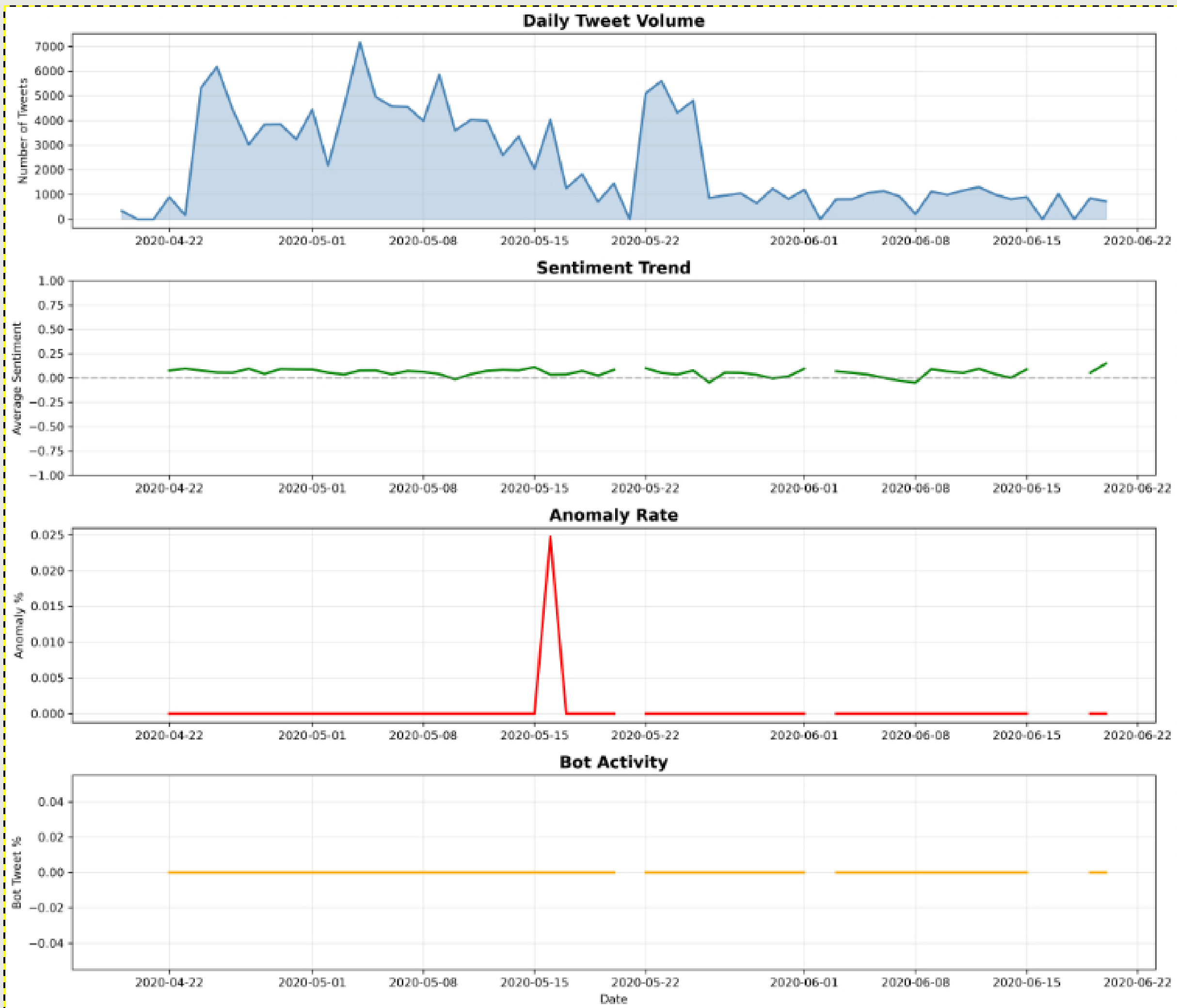

Figure 7: Trend and Topic Analysis Dashboard

The distribution of topic volumes underscores a highly uneven structure in thematic engagement. Topic ID 0 dominates the discourse with roughly 28,000 tweets, establishing it as the principal locus of attention. The remaining topics display significantly lower frequencies. The deliberate use of color coding in the visualization highlights Topics 9 and 14 in red, setting them apart from the green and gray bars assigned to the others. This visual contrast signals potential thematic importance, irregularity, or risk associated with these two topics, prompting further examination of their content and temporal patterns.

Emotional characterization across the topics reveals an unexpectedly uniform landscape. The emotion heatmap, which spans Topic IDs 0 through 14, registers a consistent 100 percent classification under the neutral category. This uniformity indicates that the current emotion detection model fails to capture affective differentiation across themes. The absence of anger, fear, sadness, or joy across all topics suggests either an inherent emotional flatness in the dataset or, more plausibly, an overly conservative or undertrained emotion classifier that requires recalibration or enrichment to detect subtler affective cues.

The topic co-occurrence network provides a complementary structural perspective on discourse organization. The visualization displays an extremely dense central cluster, overwhelmingly shaped by Topic 0, which acts as a gravitational center for the majority of topical interactions. In contrast, numerous peripheral topics appear isolated or weakly linked, forming small clusters or solitary nodes along the network's boundaries. This architecture implies a highly centralized conversational ecosystem in which one dominant theme subsumes most of the discussion, while many other identified topics remain marginal, episodic, or disconnected from the core narrative.

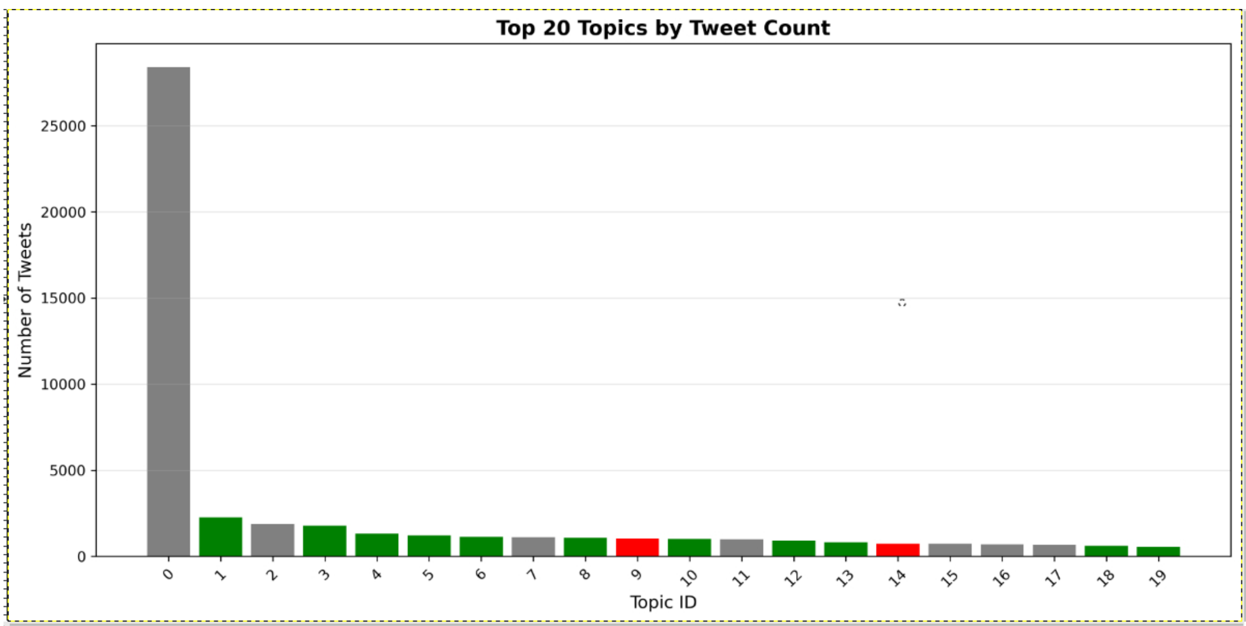

Figure 8: Top 20 Topics BY Tweet Count

Figure 8 shows the top twenty topics by tweet count. According to the figure, Topic ID 0 overwhelmingly dominates the dataset, accounting for approximately 28,000 tweets. Its volume is significantly higher than that of any other topic, indicating that a single subject drives the majority of the conversation captured. This distribution reflects a highly skewed, long-tail pattern, where a few topics accumulate the bulk of activity while the remaining topics exhibit comparatively minimal engagement.

Topic ID 1 ranks as the second most frequent topic, with roughly 2,000 tweets. This represents a steep decline from Topic 0, underscoring the asymmetry in topic prevalence within the dataset. Topics 1 through 7 each register between approximately 1,000 and 2,000 tweets, showing moderate but relatively similar engagement levels.

The remaining topics, spanning Topic IDs 8 through 19, display very low and nearly uniform volumes, generally at or below 1,000 tweets. These peripheral topics collectively highlight the long-tail nature of the distribution, where a large number of topics contribute minimally to the overall discourse compared to the dominant themes.

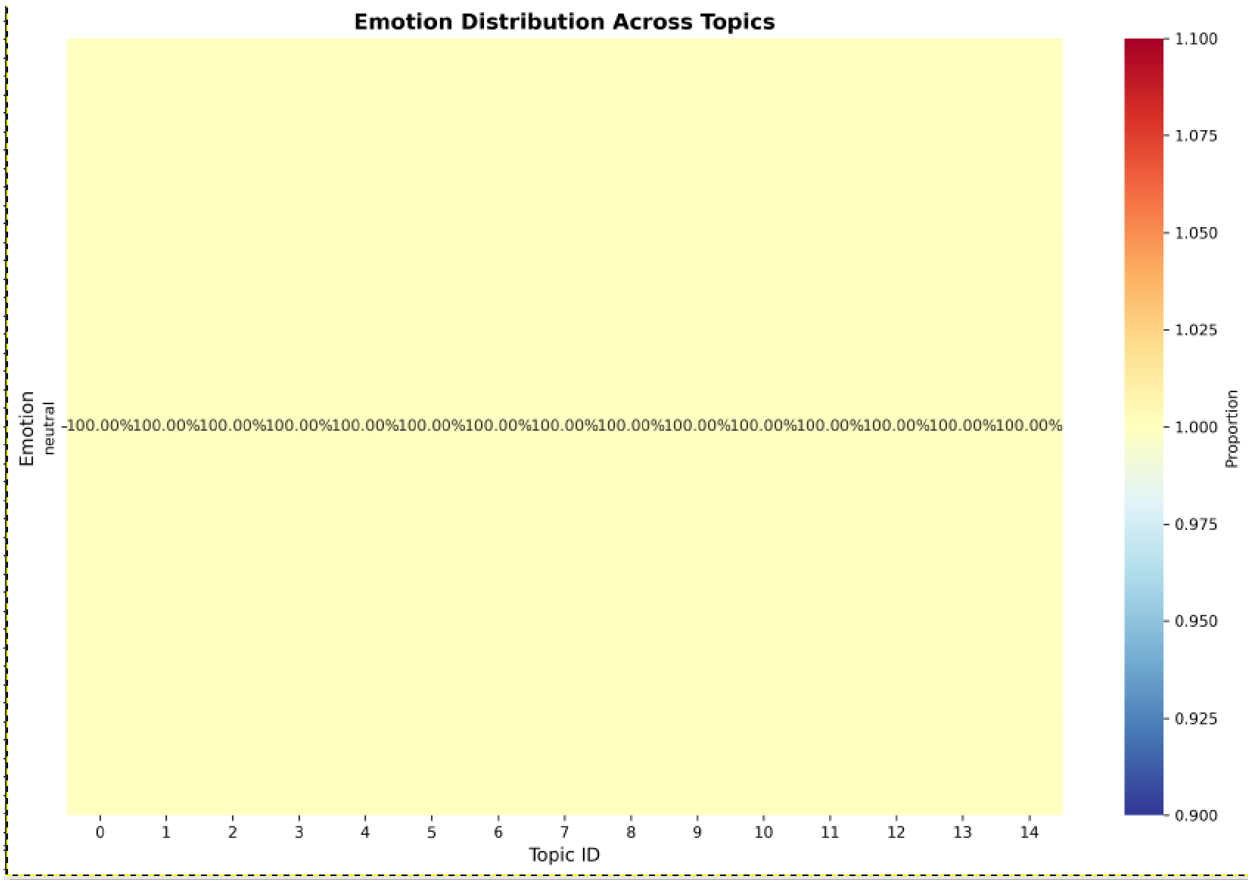

Figure 9: Sentiment Analysis Heatmap

Figure 9 shows the sentiment analysis heatmap. The heatmap shows the proportion of different emotions across topics IDs 0 through 14. Only one emotion—Neutral—is explicitly represented, providing a focused view of affective content as captured by the model.

The key finding from this visualization is striking: every topic, from Topic ID 0 to Topic ID 14, is classified as completely neutral. The heatmap displays a uniform yellow color across all topics, corresponding to a proportion of 1.000 (completely) on the color bar. This complete uniformity indicates the absence of any detected positive, negative, or other emotional signals within the top topics.

The implications of this observation are notable. It suggests that the sentiment model employed either has a strong bias toward detecting neutrality or is configured in a way that prevents recognition of non-neutral emotions.

Figure 10 shows the Topic Co-occurrence Network. The topic co-occurrence network visualizes the frequency with which different topics appear together within individual social media posts. The graph reveals a highly centralized structure, with a massive purple cluster dominating the visualization. At the core of this cluster lies Topic 0, reinforcing findings from the “Top 20 Topics by Tweet Count” that identify it as the most dominant and central topic in the dataset.

The density of the central cluster indicates high interconnection among Topic 0 and several other top-ranking topics, likely including Topics 1, 2, and 3. This pattern suggests that posts addressing Topic 0 frequently reference . It forms a cohesive and highly integrated core conversation.

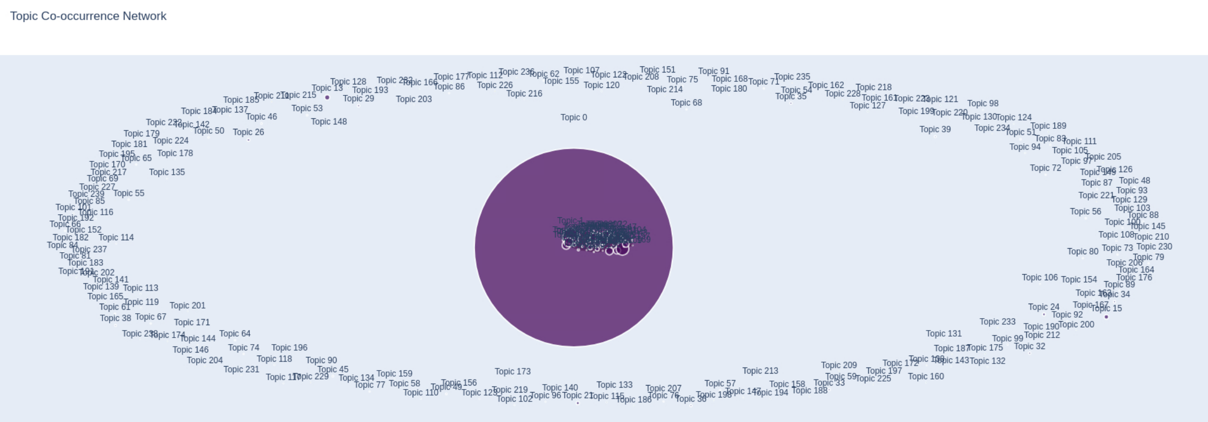

Figure 10: Topic Co-occurrence Network

A large number of peripheral topics, such as Topics 128, 196, and 209, are positioned along the network’s edges with minimal connections to the central cluster. These peripheral nodes likely represent niche or isolated discussions, topics that are rarely mentioned in conjunction with the main conversation.

4.2 Discussion, Lessons Learnt and Future Directions for the first run

The results show that the topic modeling process was largely unrefined. The clustering algorithm has failed to separate generic linguistic noise from specific thematic signals. Topic 0 seems to function as a garbage bin for common stop words or platform-specific filler. it creates a statistical vacuum that pulls in the majority of the data. This dominance forces a scale compression on the Y-axis that effectively hides the nuanced variance between the remaining 19 topics, making a meaningful comparison between them difficult without removing the primary outlier.

The strategic application of color introduces a qualitative layer that overrides the quantitative hierarchy. While Green represents the dominant "healthy" discourse and Grey represents neutral or unclassified data, the Red bars (Topics 9 and 14) serve as critical flags. Structurally, these are not volume leaders, but their color-coding suggests they are "Topic Outliers" based on content risk or sentiment rather than tweet count. This creates a dual-layered analysis where the viewer must balance the massive physical presence of the grey/green bars against the high-priority, low-volume "threat" or "anomaly" signaled by the red indicators.

The nearly flat "tail" from Topic 1 through Topic 19 indicates a fragmented and decentralized narrative environment. In a high-impact event, one would expect a "Stepped Decay" where several large sub-narratives emerge. Instead, once the "noise" of Topic 0 is removed, the remaining data reflects a sea of micro-conversations with no clear secondary dominance. This suggests that the dataset captures a broad, unfocused period of time where no single news event or viral trend was strong enough to consolidate the discourse into distinct, high-volume clusters.

The temporal activity analysis highlights that while tweet volumes vary substantially, the models fail to detect nuanced changes in sentiment, anomalies, or bot activity. The sentiment trend remains slightly positive and nearly flat, the anomaly rate is essentially zero except for a single spike, and bot activity is underestimated. These shortcomings indicate that the current models lack sensitivity. Future work should focus on fine-grained detection because accurately capturing fluctuations in sentiment, anomalies, and automated behavior is essential for understanding the dynamics and authenticity of social media discourse. Without this sensitivity, interpretations of user behavior or engagement patterns may be misleading, limiting the usefulness of the analysis for real-world applications.

At the topic modeling level, Topic ID 0 dominates overwhelmingly, with other topics showing steeply diminishing frequencies. Many topics are unnamed, and the model misrepresents the distribution of conversation. Addressing this is crucial because unbalanced or poorly labeled topics reduce interpretability, obscure minority discussions, and prevent stakeholders from understanding the diversity of themes. Future improvements should include reducing the number of topics to a manageable set, ensuring coverage is balanced across topics, and assigning meaningful, real-world labels. This will enhance clarity, allow for actionable insights, and support human-in-the-loop validation of model outputs.

Emotion detection also shows clear limitations, as all topics are classified as completely neutral. This is concerning because ignoring subtle emotional signals limits the ability to understand the affective tone of discourse and may result in misinformed conclusions about public sentiment. Future work should expand emotion categories beyond neutrality, improve model sensitivity to detect positive, negative, and mixed emotions, and validate against human-labeled benchmarks. Such refinements are critical to ensuring that the model captures the true emotional complexity of social media content.

The topic co-occurrence network further emphasizes the need for careful modeling. Topic 0 dominates the central cluster, while peripheral topics are weakly connected or isolated. This indicates that the current approach may artificially exaggerate the prominence of central topics and underestimate the independence of niche discussions. Future directions should focus on improving granularity, incorporating confidence scores, and labeling topics clearly. These measures will help ensure that network representations accurately reflect genuine conversational relationships, supporting reliable interpretation of discourse patterns.

The code requires more visualizations, including representations of confidence scores, which are essential for assessing the reliability and success of the model. Incorporating these visual outputs will improve interpretability and enable more transparent evaluation of model performance.

4.3 Results for the second run

Figure 11 shows the model confidence across various components. The figure reveals a pronounced imbalance in model confidence across tasks. Language processing achieves uniformly high confidence, reaching a perfect average score, which indicates that the model performs this task with complete certainty across the dataset. In contrast, all analytical tasks that require semantic, affective, or inferential reasoning exhibit low confidence. Topic and Sentiment confidence scores remain close to 0.30, while Detection is even lower, around 0.26. Emotion analysis attains a higher value at 0.50, yet this still falls below the medium confidence threshold. Collectively, these values indicate that the model struggles to assert reliable judgments beyond surface level linguistic identification.

The distributional characteristics of the confidence scores further clarify the nature of this under-performance. Topic confidence is consistently low, with most predictions clustered tightly around the mean. This indicates systematic uncertainty rather than occasional failure, suggesting that the model rarely encounters instances it can classify with strong confidence. Sentiment confidence, by contrast, is highly variable. The model alternates between very low and very high confidence assignments, pointing to unstable or poorly calibrated sentiment representations. Emotion confidence shows the opposite behavior. Scores concentrate almost entirely around a single mid range value, implying that the model defaults to a neutral confidence irrespective of input complexity or clarity.

Temporal analysis of the confidence scores shows no meaningful change over the observed period. Confidence levels for Topic, Sentiment, and Emotion remain flat and stable, with no upward or downward trends. This stability indicates that the low confidence is not driven by transient noise or short term shifts in the data. Instead, it reflects a persistent performance ceiling. The model behaves consistently but remains consistently uncertain in these tasks.

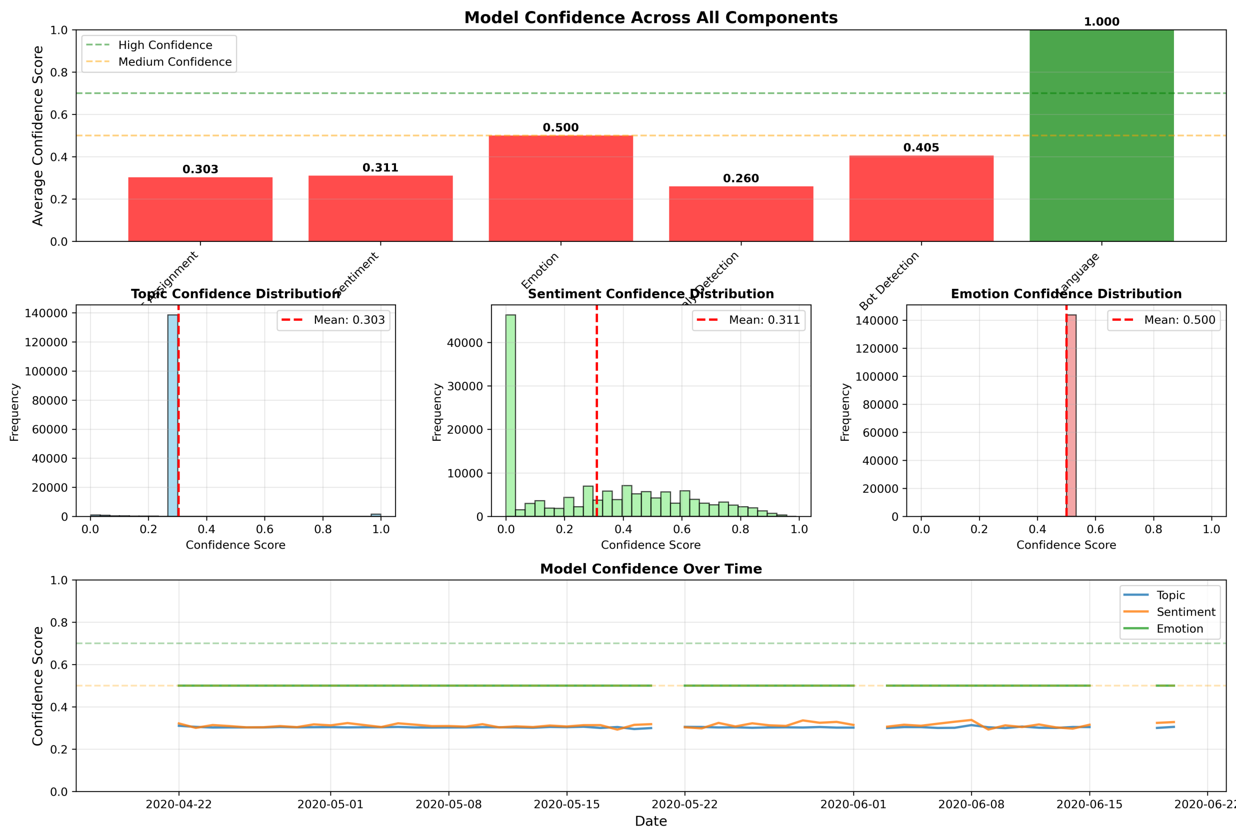

Figure 11: Topic Co-occurrence Network

Figure 12 indicates a highly volatile pattern of user engagement over the observed period. Tweet volume fluctuates sharply, with multiple surges reaching several thousand tweets per day, followed by pronounced declines. These fluctuations suggest episodic attention driven by external events rather than sustained discussion. Periods of low volume later in the timeline indicate a waning public engagement or a shift of attention away from the topic. The presence of data gaps further implies discontinuities in collection or relevance, reinforcing the episodic nature of discourse during this period.

Sentiment outcomes remain remarkably stable despite these large volumes. Average sentiment stays close to neutral and only marginally positive throughout, never approaching strongly polarized values. Variability around the mean is also consistent, indicating that while individual opinions may differ, the overall emotional tone of the discourse does not experience shocks or directional shifts. This stability suggests that even during high attention periods, public reaction remains balanced rather than emotionally extreme.

In contrast, the model's confidence in its sentiment predictions is persistently low. Confidence values consistently hover around 0.30 to 0.35 and never approach the defined adequacy threshold. This sustained low confidence mirrors earlier aggregate findings and confirms that uncertainty is not episodic or data dependent but systemic. Even when sentiment itself is stable and near neutral, the model does not express certainty in its assessments, pointing to limitations in calibration or representational strength rather than volatility in the underlying signal.

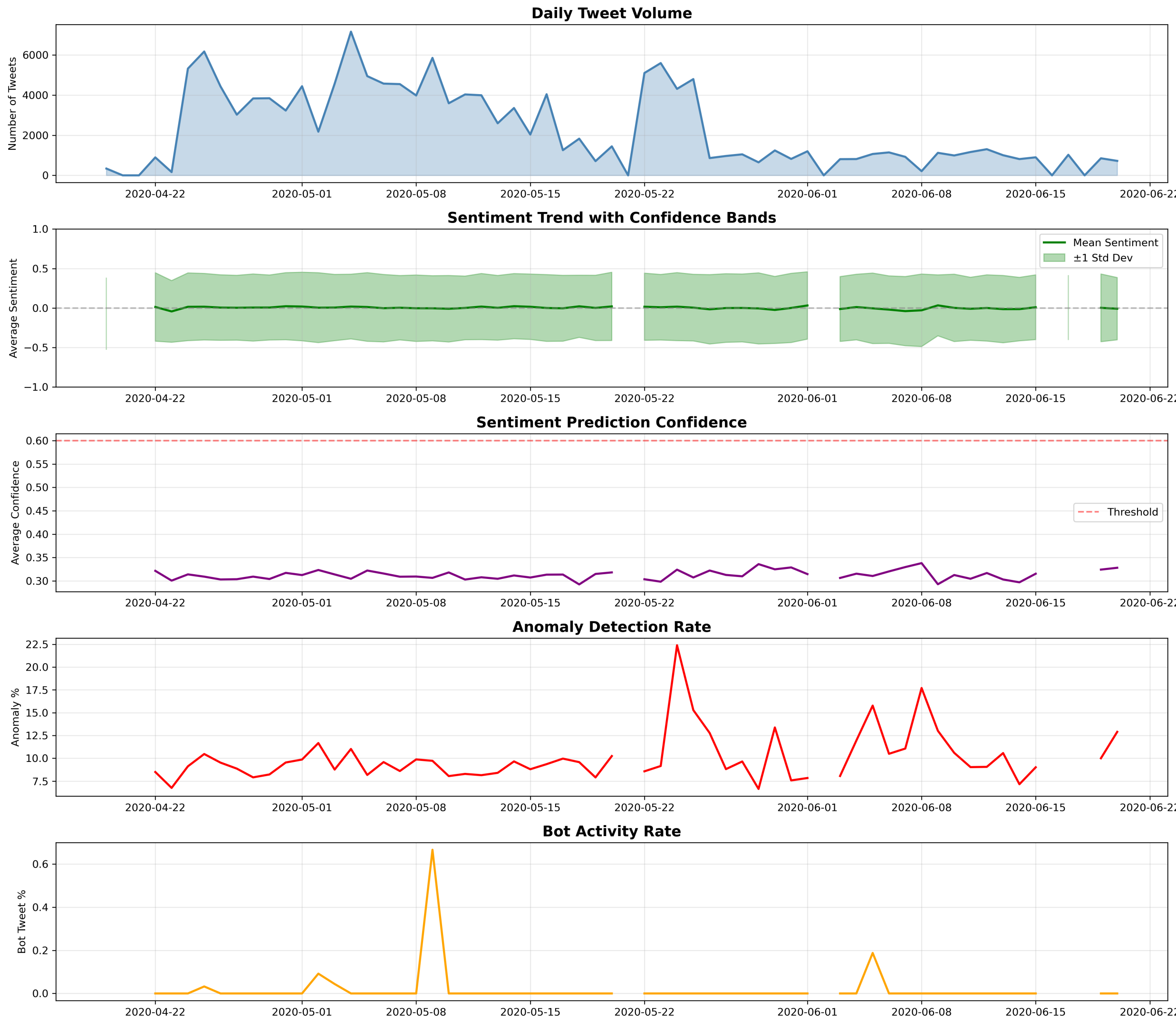

Figure 12: Temporal Trends

Anomaly detection results reveal sharp, event driven deviations in content patterns. While the baseline anomaly rate remains moderate, several spikes stand out clearly. The most pronounced spike coincides with a major increase in tweet volume, indicating that the surge was driven by substantively different or unexpected content rather than routine amplification. Additional anomaly spikes occur independently of volume peaks, suggesting discrete moments when discourse structure or semantics diverged from learned norms.

Bot activity remains negligible overall, with rates close to zero for most of the period. However, isolated spikes show brief but concentrated bursts of automated behavior. These spikes do not align with either volume surges or anomaly peaks, indicating that bot activity is not the primary driver of large scale engagement or anomalous content. Instead, bot behavior appears sporadic and largely decoupled from broader discourse dynamics.

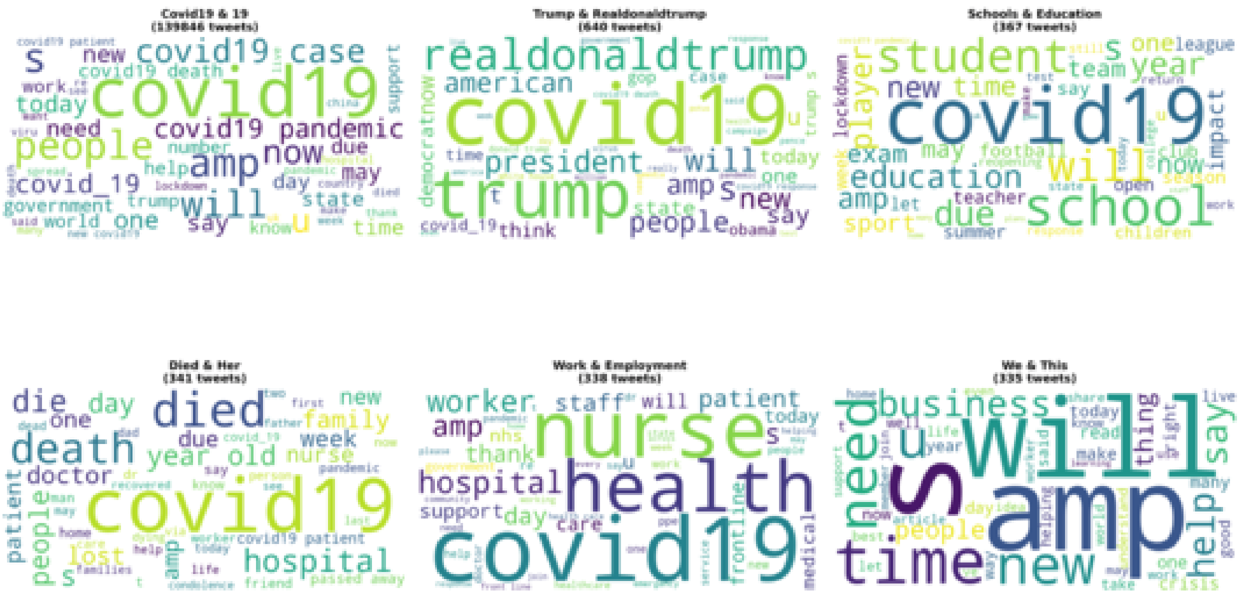

Figure 13: Temporal Trends

Figure 13 shows the topic word cloud. This is dominated by the term COVID-19. For future iterations, this topic must be excluded as it is the overall topic of the entire dataset and it overshadows other topics.

Figure 14 shows the UMAP projections. The results show that the underlying discourse occupies a single, dense, and highly interconnected semantic space. Tweets are not organized into clearly separable groups but instead form a continuous cloud of closely related content. This indicates that the conversation is thematically cohesive at a high level, with substantial overlap in language, concepts, and framing across posts. As a consequence, distinctions between analytical categories emerge only weakly in the learned representation.

Topic assignment results reflect this overlap directly. A small number of topics dominate the dataset, while others occur sparsely and without forming distinct semantic regions. Topic boundaries are diffuse rather than discrete, suggesting that the features distinguishing one topic from another are subtle and frequently shared. This lack of orthogonality implies that topic labels capture gradations of discussion rather than independent themes, which explains the persistently low confidence associated with topic prediction. The model is not failing sporadically but is operating in a space where strong topic separability does not exist.

Sentiment results exhibit a similar pattern. Neutral sentiment overwhelmingly dominates, with positive and negative expressions embedded within the same semantic neighborhoods. Sentiment does not reorganize the data into distinct regions, indicating that emotional polarity is layered onto shared content rather than defining separate modes of discourse. Slight tendencies for positive and negative expressions to appear toward different edges of the space exist, but these effects are weak and insufficient to support high confidence classification. This structural entanglement accounts for both the near neutral average sentiment and the model's chronic uncertainty in sentiment prediction.

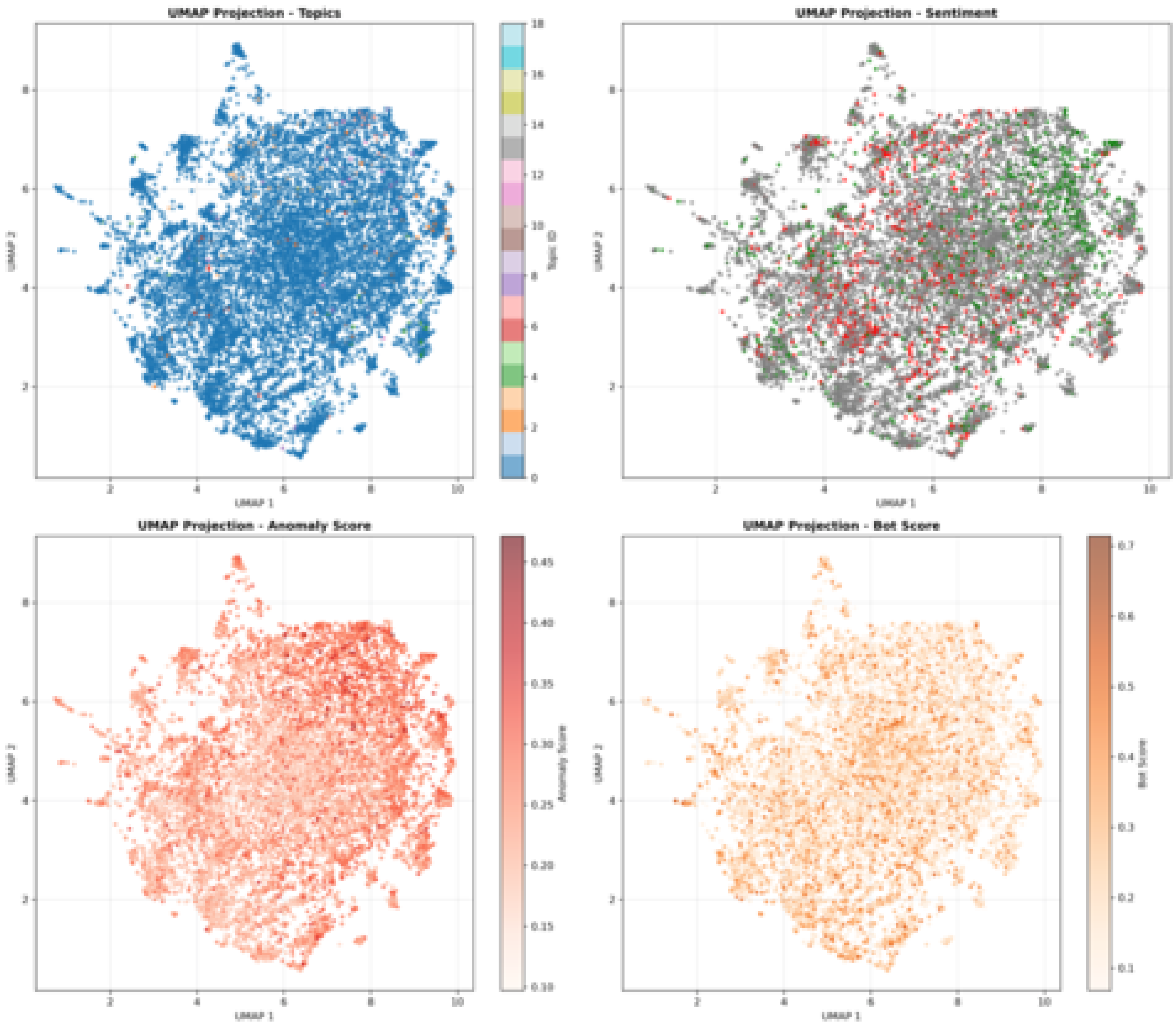

Figure 14: UMAP Projections for the four tasks

Anomalous content is revealed to be embedded within the core of the conversation rather than isolated at its periphery. Tweets with higher anomaly scores are not structurally distant from typical posts, but instead resemble common discourse while containing unusual combinations of features or unexpected content signals. This indicates that anomalies arise from semantic deviation within familiar conversational patterns rather than from entirely novel or foreign structures. As a result, anomaly detection highlights moments of qualitative change in what is being said, not breaks in how discourse is organized.

Bot related results stand in sharp contrast to the content driven metrics. Tweets with elevated bot likelihood are extremely rare and do not occupy any coherent region of the semantic space. Their dispersion indicates that bot characteristics are largely independent of topical, sentimental, or structural features of the text. Automated behavior manifests through external account level signals rather than through distinctive linguistic patterns that would reposition tweets within the shared semantic manifold.

The UMAP projections collectively demonstrate that the model is operating on a highly intertwined and complex dataset where topic, sentiment, and anomaly are not clearly separable in the low-dimensional feature space. This high degree of feature overlap confirms and visually explains the consistently low confidence scores observed for Topic and Sentiment prediction in the time series data. Bot activity, conversely, is an orthogonal characteristic, confirming it's driven by external account features rather than the content's structural relationship to other tweets.

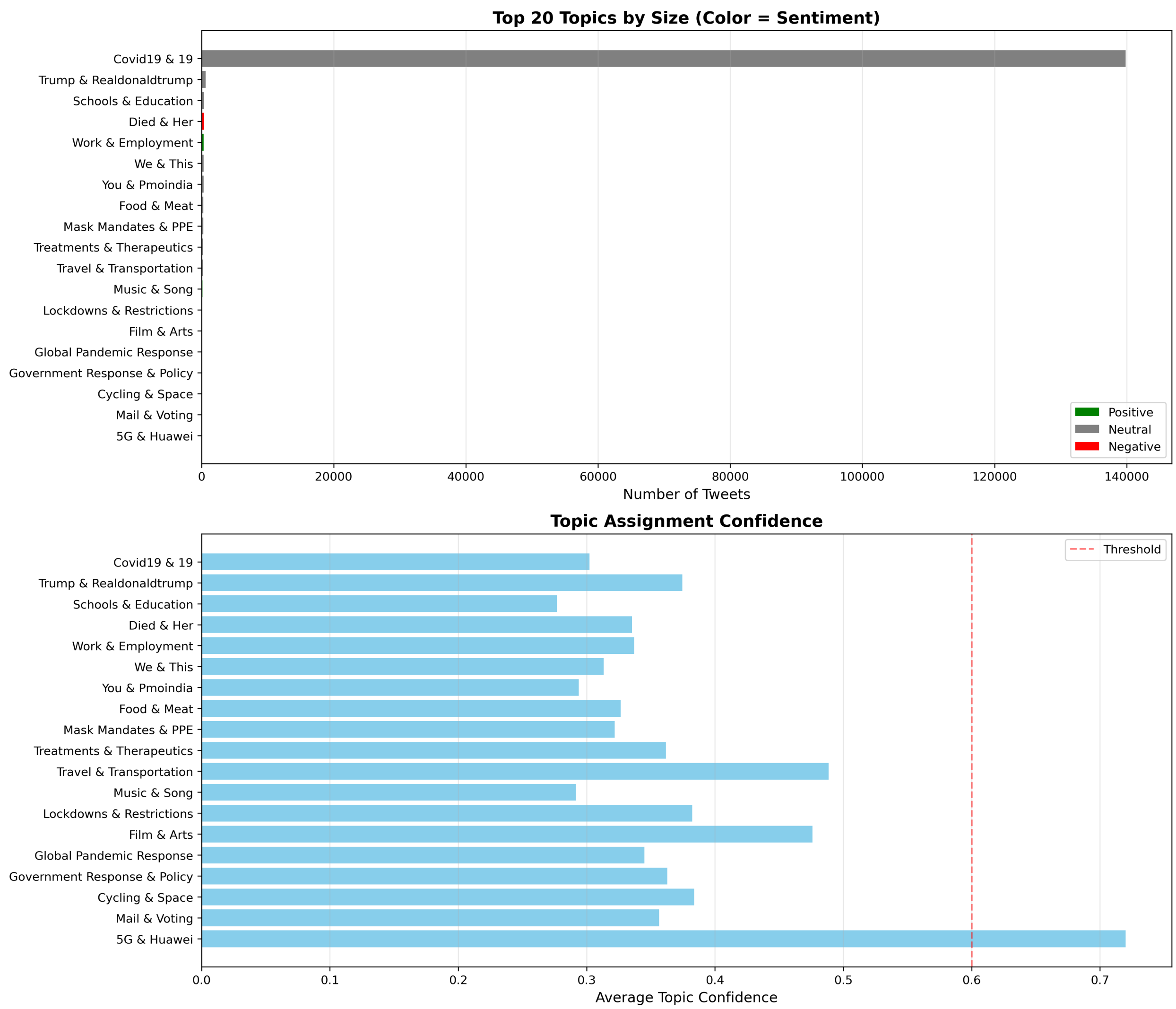

Figure 15: Topic Distribution

The results from topic distribution (Figure 15) reveal an extreme imbalance in topical distribution. One topic overwhelmingly dominates the dataset, accounting for the vast majority of tweets, while all remaining topics contribute only marginally. This concentration indicates that the discourse is largely centralized around a single thematic axis, with other topics functioning as peripheral or episodic subthreads rather than independent streams of discussion. Such skewness suggests that the conversation is driven by a unifying macro narrative rather than a diverse set of equally salient themes.

Sentiment associated with topic volume is largely neutral. The dominant topic exhibits a predominantly neutral emotional tone, indicating widespread informational exchange rather than polarized or emotionally charged reactions. Most smaller topics also display neutral sentiment, reinforcing the observation that the overall discourse is characterized by restraint and balance. A small number of low volume topics deviate from this pattern, showing clearer negative or positive sentiment. However, these emotionally polarized topics remain minor in scale and do not substantially influence the aggregate sentiment profile of the dataset.

Topic assignment confidence varies considerably across themes, but remains low for most. Only a small subset of topics achieves confidence levels near or above the defined adequacy threshold, suggesting that these themes are more semantically coherent and easier for the model to identify. In contrast, the majority of topics, including the most frequent one, exhibit average confidence values clustered in the low to mid range. This indicates that many tweets are assigned to topics with only moderate certainty, reflecting overlapping vocabularies and blurred thematic boundaries.

Importantly, topic size does not correspond to topic clarity. The largest topic does not achieve high assignment confidence, demonstrating that frequency alone does not imply semantic distinctiveness. Instead, large topics appear to function as broad catch all categories that absorb heterogeneous content. Smaller topics with higher confidence likely represent more narrowly defined narratives or specific entities, allowing the model to assign tweets with greater certainty despite their limited volume.

This further suggests that COVID-19 must be excluded from the list of topics in future iterations.

5 Discussion

A set of interconnected limitations in the model are revealed. These limitations go beyond isolated performance issues and instead point to structural mismatches between the modeling assumptions and the nature of COVID-19 Twitter discourse. These problems manifest consistently across topic size, confidence scores, sentiment behavior, anomaly patterns, and low-dimensional embeddings, indicating that the model is struggling not because of noise or data scarcity, but because of how it conceptualizes and represents meaning in this domain.

A primary problem is the model's reliance on discrete categorical assignments in a fundamentally continuous and overlapping semantic space. Topic modeling assumes that tweets can be reasonably mapped to separable thematic clusters. However, COVID-19 discourse is inherently entangled. Individual tweets often blend health updates, policy reactions, personal experiences, and political commentary in a single sentence. The results show that the model responds to this by collapsing most content into one dominant topic while assigning the rest with low confidence. This is not merely topic imbalance; it is evidence that the learned topic representations lack sufficient discriminatory power to separate conceptually adjacent narratives. The dominant topic becomes a semantic catch-all, absorbing ambiguity rather than resolving it.

The low topic assignment confidence across nearly all topics, including the largest one, further highlights this issue. Confidence does not increase with volume, which suggests that the model is not learning sharper boundaries as it sees more examples. Instead, increased exposure reinforces ambiguity. This indicates that the latent topic space is poorly aligned with the way pandemic discourse is structured. Topics appear to be defined by shallow lexical cues rather than by deeper contextual or relational semantics, causing the model to hesitate when multiple cues coexist.

Sentiment analysis reveals a parallel but distinct failure mode. Average sentiment remains neutral and stable, yet the model consistently reports low confidence in its sentiment predictions. This combination suggests that neutrality is often a default outcome rather than a confident judgment. The model appears unable to distinguish between genuinely neutral statements and emotionally charged statements expressed in restrained or informational language. In the context of COVID-19, where fear, frustration, and concern are frequently conveyed implicitly, sentiment is often encoded pragmatically rather than lexically. The model's uncertainty indicates that it is insufficiently sensitive to these implicit emotional signals and overly dependent on explicit sentiment markers.

The confidence distributions and their stability over time reinforce the interpretation that these are not transient calibration problems. Confidence does not improve during high-volume periods or degrade during low-volume ones. This temporal flatness suggests that the model has reached a performance ceiling imposed by its representational capacity or training paradigm. In other words, the model is consistently uncertain because it lacks the expressive machinery needed to resolve ambiguity in short, context-poor texts discussing a complex, evolving crisis.

The UMAP-based results provide a geometric explanation for these failures. Tweets occupy a dense, continuous manifold with no clear separations by topic or sentiment. The model attempts to impose discrete labels onto a space where variation is gradual and multidimensional. Anomalous tweets are embedded within the core of the discourse rather than isolated, indicating that novelty arises from subtle semantic shifts rather than structural deviation. This further challenges a model that expects anomalies to be outliers and topics to form islands.

Another important limitation is the model's asymmetric competence across tasks. Language identification achieves perfect confidence, while higher-order tasks such as topic, sentiment, and emotion remain weak. This asymmetry reveals that the model is effective at surface-level pattern recognition but struggles with interpretive reasoning. It can reliably identify what language is used, but not what is being said or felt in a nuanced sense. This gap suggests overreliance on pretrained linguistic features without sufficient task-specific adaptation for crisis discourse.

Finally, the contrast with bot detection is revealing. Bot activity is sparse and cleanly identified, largely independent of content semantics. This success highlights that the model performs best when the task relies on orthogonal, well-defined signals. Where signals are entangled, implicit, and context dependent, performance degrades sharply. This contrast underscores that the model's core weakness lies in modeling human meaning under uncertainty, not in handling structured or externally defined features.

6 Future Directions

The future course should focus on reframing the modeling strategy to align with the nature of COVID-19 Twitter discourse rather than relying on incremental adjustments. The current model is stable but misaligned with the complex, overlapping, and nuanced structure of the data.

One approach is to move away from rigid, discrete topic assignments and adopt soft, multi-topic representations. Tweets should be allowed to belong to multiple topics simultaneously with graded weights. This would better capture overlapping narratives and reduce the dominance of a single catch-all topic. Dynamic or hierarchical topic models could further distinguish macro themes like "COVID-19" from more specific subthemes such as policy, health outcomes, and personal experiences.

Sentiment and emotion modeling should shift toward contextual and uncertainty-aware inference instead of simple polarity classification. Rather than forcing neutral, positive, or negative labels, the model could estimate affective dimensions like concern, risk, trust, or urgency, with calibrated confidence levels. Using contextual embeddings fine-tuned on crisis-related data would improve detection of implicit or subtle emotional signals.

Confidence scores should be treated as informative outputs rather than secondary diagnostics. Low confidence should guide downstream analysis, highlighting tweets or periods for human review, active learning, or deeper qualitative assessment instead of being ignored or averaged.

Anomaly detection should focus on semantic and narrative shifts rather than isolated outlier tweets. Since unusual content is embedded within the main discourse, the model should detect changes in the trajectory of conversations over time, capturing how events alter the structure and content of discussions. Temporal clustering or change-point detection in embedding space can help identify these patterns.

The modeling architecture could benefit from a joint or multitask framework, where topic, sentiment, and anomaly detection inform each other. Learning shared representations across tasks would improve robustness and reduce fragmentation, helping the model capture deeper semantic relationships instead of relying on shallow, task-specific features.

Evaluation should prioritize faithful representation of ambiguity and overlap. Metrics should account for uncertainty, and analysis should include qualitative validation, alignment with external events, and case studies to ensure the model captures meaningful patterns rather than forcing artificial clarity.

7 Conclusions

Our work aims to establish a comprehensive framework for analyzing COVID-19-related Twitter discourse by integrating topic modeling, sentiment estimation, bot segregation, and anomaly detection within a unified pipeline. While the framework was designed to provide multidimensional insights into online discourse, the current evaluation reveals several limitations. Topic modeling disproportionately emphasizes a single topic, reducing thematic granularity, while sentiment and emotion detection largely fail to capture affective variation, classifying nearly all content as neutral. Anomaly and bot detection also show low sensitivity, and network analyses exaggerate central topic dominance while underrepresenting peripheral discussions.

These findings highlight areas for refinement, such as improving model sensitivity for sentiment, emotion, and anomaly detection, balancing and naming topics for interpretability, and incorporating confidence scores and additional visualizations to assess model reliability. Despite these limitations, the study establishes a foundational pipeline for large-scale social media analysis and underscores the importance of iterative model evaluation. Future work focusing on these improvements will enhance the framework's ability to accurately capture authentic human discourse, providing more reliable and nuanced insights into pandemic-related conversations on social media.

References

[1] Tirendi, D., and L. Gargiulo, "Measuring the Impacts of Social Media on Daily Life from Millennials to Generation Alpha," *International Journal of Research in Humanities and So-*

cial Studies, vol. 11, no. 3, pp. 26–33, 2024. Available: https://doi.org/10.62557/2394-6296.110304

[2] DataReportal, "Digital Around the World," Kepios, Apr. 2025. Available: https://datareportal.com/reports/digital-2025-global-overview-report

[3] Y. Wang, S. Gao, and W. Gao, "Investigating dynamic relations between factual information and misinformation: Empirical studies of tweets related to prevention measures during COVID-19," *Journal of Contingencies and Crisis Management*, 2021. Available: https://doi.org/10.1111/1468-5973.12385

[4] L. Li, A. Sullivan, A. Musah, K. Stavrianaki, C. E. Wood, P. Baker, and P. Kostkova, "Positive and negative emotions during the COVID-19 pandemic: A longitudinal survey study of the UK population," *PLOS ONE*, vol. 19, no. 2, pp. e0297214–e0297214, 2024. Available: https://doi.org/10.1371/journal.pone.0297214

[5] K. Lanyi, R. Green, D. Craig, and C. Marshall, "COVID-19 vaccine hesitancy: Analysing Twitter to identify barriers to vaccination in a low uptake region of the UK," *Frontiers in Digital Health*, vol. 3, 2022. Available: https://doi.org/10.3389/fdgth.2021.804855

[6] K. Terry, F. Yang, Q. Yao, and C. Liu, "The role of social media in public health crises caused by infectious disease: A scoping review," *BMJ Global Health*, vol. 8, no. 12, p. e013515, 2023. Available: https://doi.org/10.1136/bmjgh-2023-013515

[7] S. B. Swetland, A. N. Rothrock, H. Andris, B. Davis, L. Nguyen, P. Davis, and S. G. Rothrock, "Accuracy of health-related information regarding COVID-19 on Twitter during a global pandemic," *World Medical & Health Policy*, vol. 13, no. 3, pp. 503–517, 2021. Available: https://doi.org/10.1002/wmh3.468

[8] F. Rodrigues, R. Newell, G. R. Babu, T. Chatterjee, N. K. Sandhu, and L. Gupta, "The social media infodemic of health-related misinformation and technical solutions," *Health Policy and Technology*, vol. 13, no. 2, p. 100846, 2024. Available: https://doi.org/10.1016/j.hlpt.2024.100846

[9] Y. Su, A. Venkat, Y. Yadav, L. Puglisi, and S. J. Fodeh, "Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities," *Computers in Biology and Medicine*, vol. 132, p. 104336, 2021. Available: https://doi.org/10.1016/j.compbiomed.2021.104336

[10] F. Isasi, M. D. Naylor, D. Skorton, D. Grabowski, S. Hernandez, and V. Montgomery Rice, "Patients, families, and communities COVID-19 impact assessment: Lessons learned and compelling needs," *NAM Perspectives*, vol. 11, 2021. Available: https://doi.org/10.31478/202111c

[11] N. Thakur, "Investigating and analyzing self-reporting of Long COVID on Twitter: Findings from sentiment analysis," *Applied System Innovation*, vol. 6, no. 5, p. 92, 2023. Available: https://doi.org/10.3390/asi6050092

[12] M. D. Giusta, S. Jaworska, and D. Vukadinović Greetham, "Expert communication on Twitter: Comparing economists' and scientists' social networks, topics and communicative styles," *Public Understanding of Science*, vol. 30, no. 1, pp. 75–90, 2020. Available: https://doi.org/10.1177/0963662520957252

[13] H. Y. Yan, K.-C. Yang, J. Shanahan, and F. Menczer, "Exposure to social bots amplifies perceptual biases and regulation propensity," *Scientific Reports*, vol. 13, no. 1, p. 20707, 2023. Available: https://doi.org/10.1038/s41598-023-46630-x

[14] V. Khandelwal, M. Gaur, U. Kursuncu, V. L. Shalin, and A. P. Sheth, "A domain-agnostic neurosymbolic approach for big social data analysis: Evaluating mental health sentiment on social media during COVID-19," in *Proc. 2024 IEEE Int. Conf. on Big Data (BigData)*, pp. 959–968, 2024. Available: https://doi.org/10.1109/bigdata62323.2024.10825174

[15] A. Sayed, M. M. Gomaa, and M. M. Nazier, "Sentiment analysis on Twitter's big data against the COVID-19 pandemic using machine learning algorithms," *Information Sciences Letters*, vol. 12, no. 8, pp. 2747–2756, 2023. Available: https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=2124&context=isl

[16] A. Vohra and R. Garg, "Deep learning based sentiment analysis of public perception of working from home through tweets," *Journal of Intelligent Information Systems*, 2022. Available: https://doi.org/10.1007/s10844-022-00736-2

[17] S. A. Waheeb, N. A. Khan, and X. Shang, "Topic modeling and sentiment analysis of online education in the COVID-19 era using social networks based datasets," *Electronics*, vol. 11, no. 5, p. 715, 2022. Available: https://doi.org/10.3390/electronics11050715

[18] S. D. Sosun, B. Tayfun, Y. Nukan, I. Altun, E. B. Erik, E. Yildirim, B. Kocacinar, and F. P. Akbulut, "Deep sentiment analysis with data augmentation in distance education during the pandemic," in *Proc. 2022 Innovations in Intelligent Systems and Applications Conf. (ASYU)*, pp. 1–5, 2022. Available: https://doi.org/10.1109/asyu56188.2022.9925379

[19] F. Azizi, H. VahdatNejad, and H. Hajiabadi, "Temporal analysis of topic modeling output by machine learning techniques," *International Journal of Data Science and Analytics*, vol. 20, no. 3, pp. 2175–2225, 2025.

[20] A. Ünlü, S. Truong, N. Sawhney, and T. Tammi, "Setting the misinformation agenda: Modeling COVID-19 narratives in Twitter communities," *New Media & Society*, 2025.

[21] E. Hashmi, A. Altaf, M. W. Anwar, M. H. Jamal, and U. I. Bajwa, "BI-SENT: Bilingual aspect-based sentiment analysis of COVID-19 tweets in Urdu language," *PLOS ONE*, vol. 20, no. 6, 2025.

[22] K. D. S. Balasooriya, R. A. H. M. Rupasingha, and B. T. G. S. Kumara, "Analysis of suicide-related tweets during the COVID-19 pandemic," *Coronaviruses*, vol. 6, no. 1, 2025.

[23] V. V. Kumar, A. Sahoo, S. K. Balasubramanian, and S. Gholston, "Mitigating healthcare supply chain challenges under disaster conditions: A holistic AI-based analysis of social media data," *International Journal of Production Research*, vol. 63, no. 2, pp. 779–797, 2025.

[24] C. Cochis, F. Bertolotti, and P. Ungureanu, "The dynamic response of coworking communities: unpacking resilience in times of disruption," *Journal of Management Inquiry*, 2025.

[25] K. A. Momin, H. M. I. Kays, and A. M. Sadri, "Identifying Crisis Response Communities in Online Social Networks for Compound Disasters: The Case of Hurricane Laura and COVID-19," *International Journal of Disaster Risk Reduction*, vol. 96, 2024.

[26] H. Gomez-Adorno, G. Bel-Enguix, G. Sierra, J.-C. Barajas, and W. Álvarez, "Machine Learning and Deep Learning Sentiment Analysis Models: Case Study on the SENT-COVID Corpus of Tweets in Mexican Spanish," *Applied Sciences*, vol. 14, no. 3, 2024.

[27] M. Bhardwaj, P. Mishra, S. Badhani, and S. K. Muttoo, "Sentiment analysis and topic modeling of COVID-19 tweets of India," *Social Network Analysis and Mining*, vol. 14, no. 1, 2024.

[28] A. Hashem Sweidan, N. El-Bendary, and E. Elhariri, "Autoregressive Feature Extraction with Topic Modeling for Aspect-based Sentiment Analysis of Arabic as a Low-resource Language," *Information Processing & Management*, 2024.

[29] R. Sadigov, E. Yıldırım, B. Kocaçınar, F. P. Akbulut, and C. Catal, "Deep learning-based user experience evaluation in distance learning," *Education and Information Technologies*, 2024.

[30] D. Badawi, "Intelligent Recommendations Based on COVID-19 Related Twitter Sentiment Analysis and Fake Tweet Detection in Apache Spark Environment," *IEEE Access*, vol. 12, 2024.

[31] U. Shahzad, T. Sengupta, A. Rao, and G. D. Sharma, "Energy in the backseat? Investigating decarbonization dialogue in supply chain tweets during and after COVID-19," *Energy Reports*, vol. 10, pp. 1186–1198, 2024.

[32] M. K. Alhumayani and H. N. Alhazmi, "Detecting Reported Side Effects of COVID-19 Vaccines From Arabic Twitter (X) Data," *Computers in Biology and Medicine*, 2024.

[33] J. Jagarapu, M. I. Diaz, C. U. Lehmann, and R. J. Medford, "Twitter discussions on breastfeeding during the COVID-19 pandemic," *International Breastfeeding Journal*, 2023.

[34] B. Mathayomchan, V. Taecharungroj, and W. Wattanacharoensil, "Evolution of COVID-19 tweets about Southeast Asian Countries: topic modelling and sentiment analyses," *Social Sciences & Humanities Open*, vol. 7, 2023.

[35] H. Yadav and M. Sagar, "Exploring COVID-19 vaccine hesitancy and behavioral themes using social media big-data: a text mining approach," *Data in Brief*, 2023.

[36] F. Azizi, H. Hajiabadi, H. Vahdat-Nejad, and M. H. Khosravi, "Detecting and analyzing topics of massive COVID-19 related tweets for various countries," *Journal of Information Science*, vol. 49, no. 5, pp. 812–826, 2023.

[37] Z. Zaidi, M. Ye, F. Samon, J. Evans, and Y. Kashima, "Topics in Antivax and Provax Discourse: Yearlong Synoptic Study of COVID-19 Vaccine Tweets," *Social Media + Society*, vol. 9, no. 3, 2023.

[38] S. Hussain, M. Ayoub, Y. Yu, D. P. F. Moller, and H. Weiyan, "Ensemble Deep Learning Framework for Situational Aspects-Based Annotation and Classification of International Student's Tweets during COVID-19," *IEEE Access*, vol. 11, pp. 32214–32227, 2023.

[39] H. Alhazmi, "Detection of Students' Problems in Distance Education Using Topic Modeling and Machine Learning," *Computers and Education: Artificial Intelligence*, vol. 3, 2022.

[40] A. Ouassini, M. Amini, and N. Ouassini, "#ChinaMustExplain: Global Tweets, COVID-19, and Anti-Black Racism in China," *Ethnic and Racial Studies*, vol. 45, no. 14, pp. 2643–2665, 2022.

[41] C. S. Lakshmi, S. Saxena, and B. S. Kumar, "Text Mining Data Summarization Approach for COVID-19 Using Hybrid Machine Learning Techniques," *International Journal of Data Science and Analytics*, 2025.

[42] K. Hussain, M. Azhar, B. Lee, M. Affan, and S. U. Khan, "ASAnalyzer: Attention based Sentiment Analyzer for Real-world Sentiment Analysis," *Neural Computing and Applications*, vol. 35, 2023.

[43] M. BenSassi, M. Abbes, and F. Atigui, "Fuzzy Based Text Quality Assessment for Sentiment Analysis," *Expert Systems with Applications*, 2023.

[44] A. Bari, M. Heymann, R. J. Cohen, M. Dilorenzo, and M. Coffee, "Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and NLP Algorithms," *JMIR Public Health and Surveillance*, vol. 8, no. 7, e34567, 2022.

[45] M. Mahdikhani, "Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of the COVID-19 pandemic," *International Journal of Information Management*, 2022.

[46] W. S. Albaldawi, R. M. Almuttairi, and M. E. Manaa, "Big Data Analysis for Healthcare Application using Minhash and Machine Learning in Apache Spark Framework," *Health Informatics Journal*, 2022.

[47] A. O. Alqabbany and A. M. Azmi, "Measuring the effectiveness of adaptive random forest for handling concept drift in big data streams," *Applied Soft Computing*, vol. 111, 2021.

[48] L. M. Muñoz, M. F. Ramirez, and J. E. Camargo, "A Data-Driven Method for Measuring the Negative Impact of Sentiment Towards China in the Context of COVID-19," *Asian Journal of Communication*, 2020.

[49] F. Binsar and T. Mauritsius, "Mining of Social Media on COVID-19 Big Data Infodemic in Indonesia," *International Journal of Advanced Computer Science*, 2020.

[50] X. Zhang, H. Saleh, E. M. G. Younis, R. Sahal, and A. A. Ali, "Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System," *Future Generation Computer Systems*, 2020.

[51] F. K. Sufi, "Automatic identification and explanation of root causes on COVID-19 index anomalies," *IEEE Access*, vol. 11, 2023.

[52] S. Kumar, M. B. Khan, M. H. A. Hasanat, A. AlTameem, and M. AlKhathami, "An Anomaly Detection Framework for Twitter Data," *Journal of Information and Telecommunication*, vol. 6, 2022.

[53] A. Ünlü, S. Truong, N. Sawhney, J. Sivelä, and T. Tammi, "Tracing the dynamics of misinformation and vaccine stance in Finland amid COVID-19," *New Media & Society*, 2025.

[54] S. Dadkhah, X. Zhang, A. G. Weismann, A. Firouzi, and A. A. Ghorbani, "The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker," *Data in Brief*, 2024.

[55] A. Ünlü, S. Truong, N. Sawhney, J. Sivelä, and T. Tammi, "Long-term assessment of social amplification of risk during COVID-19: challenges to public health agencies amid misinformation and vaccine stance," *Information Processing & Management*, 2024.

[56] M. Antenore, J. M. Camacho Rodriguez, and E. Panizzi, “A Comparative Study of Bot Detection Techniques With an Application in Twitter COVID-19 Discourse,” *Computers & Security*, vol. 133, 2023.

[57] H. R. Greve, H. Rao, P. Vicinanza, and E. Y. Zhou, “Online Conspiracy Groups: Micro-Bloggers, Bots, and Coronavirus Conspiracy Talk on Twitter,” *American Behavioral Scientist*, vol. 66, no. 12, pp. 1579–1604, 2022.

[58] A. M. Russell, D. Valdez, S. C. Chiang, H.-C. Lin, and P. M. Massey, “Using Natural Language Processing to Explore “Dry January” Posts on Twitter: Longitudinal Infodemiology Study,” *JMIR Public Health and Surveillance*, 2022.

[59] J. D. Moffitt, C. King, and K. M. Carley, “Hunting Conspiracy Theories During the COVID-19 Pandemic,” *Social Media + Society*, 2021.

[60] K. A. Qureshi, R. A. S. Malick, M. Sabih, and H. Cherifi, “Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter,” *IEEE Access*, vol. 9, pp. 141702–141713, 2021.