Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BioGraph-SentiCOVID: Calibrated Adaptive Context Selection for Graph Neural Sentiment Analysis of COVID-19 Tweets

Moustafa MAASKRI
Université IBN Khaldoun Tiaret
Mohamed GOISMI
mohamed.goismi@univ-tiaret.dz

Université IBN Khaldoun Tiaret
Djamal SEGHIER
Université IBN Khaldoun Tiaret
Mohamed DEBBAB
Université IBN Khaldoun Tiaret

Article

Keywords: Sentiment analysis, COVID-19, Twitter, Graph Neural Networks, Graph Attention Networks, GATv2, Bio-inspired optimisation, Genetic Algorithm, NLP, Conversational graphs, Imbalanced learning, Hyperparameter tuning, Adaptive context selection

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9602425/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BioGraph-SentiCOVID: Calibrated Adaptive Context Selection for Graph Neural Sentiment Analysis of COVID-19 Tweets

Moustafa MAASKRI[1], Mohamed GOISMI[2*], Djamal SEGHIER[1], Mohamed DEBBAB[2]

[1]Department of computer Sciences, LRIAS Laboratory, University of Ibn Khaldoun, Tiaret, Algeria.
[2*]Department of Sciences and Technology, University of Ibn Khaldoun, Tiaret, Algeria.
[1]Department of computer Sciences, LRIAS Laboratory, University of Ibn Khaldoun, Tiaret, Algeria.
[2]Department of Sciences and Technology, University of Ibn Khaldoun, Tiaret, Algeria.

*Corresponding author(s). E-mail(s): mohamed.goismi@univ-tiaret.dz;
Contributing authors: moustafa.maaskri@univ-tiaret.dz;
djamal.seghier@univ-tiaret.dz; mohamed.debbab@univ-tiaret.dz;

Abstract

COVID-19-related discussions on Twitter have generated an unprecedented volume of public-opinion data, offering valuable insights into emotional reactions, crisis communication, and societal responses to the pandemic. Accurate sentiment classification of such content is challenging owing to short noisy tweets, conversational ambiguity, class imbalance, and the sensitivity of graph neural architectures to hyperparameter choices. This paper introduces **BioGraph-SentiCOVID**, a calibrated adaptive-context graph neural framework for COVID-19 tweet sentiment analysis. Twitter conversations are modelled as directed multi-relational graphs, where tweets form nodes and reply/root edges encode conversational structure. The main methodological contribution is *Calibrated Adaptive Context Selection* (C-ACS): a context-gating mechanism that combines temperature-scaled ambiguity estimation, neighbour reliability, and hop distance to decide how much thread context each tweet should exploit before GATv2 aggregation. Node representations are derived from BERTweet, while

class-weighted loss and selective oversampling address sentiment imbalance. A Genetic Algorithm (GA) optimises graph depth, attention heads, hidden dimension, dropout, learning rate, weight decay, balancing ratio, and the context-gate threshold. Evaluation on COVID-Emo and supervised SemEval-2017 Task 4A adaptation uses accuracy, macro-F1, weighted-F1, per-class F1, ablation studies, and Wilcoxon signed-rank tests over $\boldsymbol{n}$**=20** runs. We additionally report emotion-to-sentiment mapping sensitivity, ACS calibration diagnostics, and a lightweight GA setting for resource-constrained use. BioGraph-SentiCOVID achieves competitive results among all evaluated baselines, with a macro-F1 of **0.829** on COVID-Emo and **0.851** under the supervised SemEval-2017 adaptation protocol. C-ACS improves the fixed-depth BioGraph variant by **+0.012** macro-F1 while reducing noisy message passing from irrelevant distant neighbours.

Keywords: Sentiment analysis, COVID-19, Twitter, Graph Neural Networks, Graph Attention Networks, GATv2, Bio-inspired optimisation, Genetic Algorithm, NLP, Conversational graphs, Imbalanced learning, Hyperparameter tuning, Adaptive context selection

1 Introduction

The COVID-19 pandemic, declared a global health emergency in January 2020, triggered a concurrent *infodemic* on social media [1]. Twitter became a primary channel through which individuals expressed fear, anger, support, denial, and acceptance surrounding the crisis [2, 3], making it a rich corpus for public-health sentiment monitoring [4, 5].

Applying existing sentiment-analysis methods to this corpus reveals three compounded challenges. *(i) Contextual dependency.* A reply appearing neutral in isolation may carry strong negative sentiment only when read within its parent thread [6, 7]. For example, "Another day, another miracle cure." is sarcastic and negative inside a thread on hydroxychloroquine misinformation, yet appears mildly positive in isolation. *(ii) Context noise.* Deeper reply-thread neighbourhoods can improve interpretation, but they also introduce off-topic replies, repeated claims, and majority-class noise. A single fixed graph depth is therefore suboptimal for all tweets. *(iii) Class imbalance and hyperparameter sensitivity.* Neutral and fear-related expressions dominate real COVID-19 tweet corpora, while graph neural networks (GNNs) are sensitive to attention heads, hidden dimensions, dropout, graph depth, and balancing ratios [8–11].

This paper introduces **BioGraph-SentiCOVID**, a unified framework that jointly addresses these challenges. Its main originality lies in a **Calibrated Adaptive Context Selection** (C-ACS) mechanism: instead of forcing the same conversation depth on every tweet, C-ACS lets each node exploit only the amount of thread context it needs. Ambiguous or sarcastic replies can draw from parents and roots, whereas self-contained tweets can suppress noisy neighbours. Unlike a raw uncertainty gate, C-ACS calibrates the local ambiguity estimator using temperature scaling before it controls graph message passing, reducing the risk that over-confident predictions suppress useful context for minority or sarcastic cases.

The key contributions are:

1. **Directed multi-relational conversation graph.** Reply and root edges are encoded as distinct relation types, preserving conversational provenance without overstating heterogeneity.
2. **Calibrated Adaptive Context Selection for GATv2.** We introduce a trainable context gate that combines temperature-scaled tweet-level ambiguity, semantic neighbour reliability, and hop distance to regulate message passing before GATv2 aggregation. This changes graph depth from a global fixed choice into a per-tweet adaptive decision.
3. **Imbalance-aware learning.** Class-weighted cross-entropy combined with selective oversampling (with Gaussian jitter on BERTweet embeddings) is applied exclusively on the training split.
4. **GA-based hyperparameter optimisation.** A Genetic Algorithm with tournament selection and Latin-Hypercube-Sampling initialisation searches a mixed discrete–continuous hyperparameter vector covering architecture, training, balancing, maximum graph depth, and context-gate threshold.
5. **Rigorous validation.** We implement the exact GATv2 dynamic-attention equations [11] and validate the framework with Wilcoxon signed-rank tests on n=20 runs plus targeted ablations: fixed-depth vs. adaptive context, ACS calibration, random-gate sanity checks, emotion-to-sentiment mapping sensitivity, SemEval native vs. adaptation protocols, random-graph baseline, context-only node removal, GAT vs. GATv2, frozen vs. two-stage BERTweet fine-tuning, cost-term ablation, lightweight GA-budget sensitivity, and balancing-ratio isolation.

The paper is organised as follows. Section 2 reviews related work. Section 3 formally defines the task. Section 4 presents the framework. Section 5 describes datasets, baselines, and implementation. Section 6 reports and analyses results. Section 7 presents ablation studies and statistical analysis. Section 8 discusses findings and limitations. Section 9 concludes.

2 Related Work

2.1 Tweet Sentiment Analysis

Early approaches relied on lexicon-based and feature-engineering methods [12–14]. Recurrent networks with GloVe embeddings [15] provided the first meaningful representation-learning gains, while BERT [16] and its Twitter-specific derivative BERTweet [17] established the importance of domain-adaptive pre-training. RoBERTa [18] and COVID-Twitter-BERT [19] have since pushed performance further on pandemic-adjacent domains.

2.2 COVID-19 Social Media Analysis

Cinelli et al. [2] characterised the infodemic volume; Sharma et al. [4] showed public sentiment toward mask policies evolved with policy announcements. Sosea and

Caragea [8] released COVID-Emo, an expert-annotated multi-emotion dataset of COVID-19 English tweets with fine-grained sentiment labels, which we adopt as our primary benchmark.

2.3 Graph Neural Networks for NLP

GNNs encode relational structure that sequential models cannot capture [20, 21]. Text-GCN [22] builds word–document graphs for document classification. *Graph Convolutional Networks* [23] aggregate neighbourhood features via *normalised summation*—they do *not* compute attention weights; the notion of "static attention" applies to the original GAT [10], not GCN. DialogueGCN [24] and RGAT [25] exploit speaker and reply relations for emotion recognition in dialogue. Relational GNNs such as R-GCN [26] are another natural option for multi-relation graphs; in this work we use a lighter edge-typed GATv2 design because the graph contains only two relation types and because the GA must evaluate hundreds of candidate architectures. GATv2 [11] corrects the rank-1 attention limitation of the original GAT by reordering the linear transformation, yielding strictly more expressive *dynamic* attention.

2.4 Bio-Inspired Hyperparameter Optimisation

Random search [27] and Bayesian optimisation [28] dominate continuous search spaces. For mixed discrete–continuous spaces—as in our setting—Genetic Algorithms (GAs) [29] handle integer and continuous genes natively. Young et al. [30] demonstrated GA-driven NAS on vision benchmarks; Lorenzo et al. [31] applied PSO to continuous DNN hyperparameters. We choose GA over PSO because our hyperparameter vector contains non-trivial discrete components (layer count, head count, hidden dimension), for which PSO requires continuous relaxations that can introduce optimisation artefacts.

2.5 Imbalanced Learning

Class-weighted cross-entropy [32] and SMOTE [33] address class skew. EDA data augmentation [34] has been applied to text; we instead jitter BERTweet embeddings, avoiding the semantic incoherence of interpolating in BERT space.

2.6 Positioning

Table 1 summarises how BioGraph-SentiCOVID differs from the closest prior work.

3 Problem Formulation

Let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ be a collection of tweets. Each tweet t_i belongs to a conversation thread identified by a root tweet t_r. The reply structure induces a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where:

- $\mathcal{V}$ contains one node v_i per tweet t_i;

Table 1: Comparison with related approaches. ✓ = present; ✗ = absent.

Method	Conv. Graph	Ctx. Emb.	Bal.	Bio-HPO	COVID
BERTweet [17]	✗	✓	✗	✗	✗
Text-GCN [22]	✓	✗	✗	✗	✗
DialogueGCN [24]	✓	✓	✗	✗	✗
RGAT [25]	✓	✓	✗	✗	✗
PSO-DNN [31]	✗	✓	✗	✓	✗
BioGraph (ours)	✓	✓	✓	✓	✓

- $\mathcal{E}$ contains two types of directed edges: *reply edges* $(v_i \to v_j) \in \mathcal{E}_{\text{rep}}$ when t_i directly replies to t_j, and *root edges* $(v_i \to v_r) \in \mathcal{E}_{\text{root}}$ connecting every non-root tweet to the thread root.

Each node v_i is associated with an initial feature vector $\mathbf{x}_i \in \mathbb{R}^d$ (computed from the tweet text) and a ground-truth sentiment label $y_i \in \mathcal{C}$, where $\mathcal{C} = \{1, 2, \ldots, C\}$ is the set of sentiment classes.

Task (node classification). Given $\mathcal{G}$, feature matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$, and labels for a training subset $\mathcal{V}_{\text{train}} \subset \mathcal{V}$, learn a function $f_\theta : \mathcal{V} \to \mathcal{C}$ that correctly predicts the sentiment label of every node in a held-out test set $\mathcal{V}_{\text{test}}$. The function f_θ is a GATv2 encoder followed by an MLP classifier, with parameters θ learned jointly.

Why graph node classification and not independent tweet classification? Because the label of v_i may be contextually ambiguous without its conversational neighbourhood. Encoding $\mathcal{G}$ lets the model aggregate information from v_i's ancestors and descendants, resolving sarcasm, stance, and irony that are invisible at the isolated-tweet level.

Note on graph type. Because $\mathcal{V}$ contains only tweet nodes, we describe $\mathcal{G}$ as a *directed multi-relational* graph (two edge types, one node type) rather than a heterogeneous graph. Future extensions may add user, hashtag, or URL nodes to warrant the heterogeneous designation.

4 Methodology

Figure 1 shows the end-to-end pipeline of BioGraph-SentiCOVID, which comprises six stages: tweet preprocessing, conversation-graph construction, BERTweet node embedding, Calibrated Adaptive Context Selection (C-ACS), GA hyperparameter optimisation, and GATv2 classification.

4.1 Tweet Preprocessing

Raw tweets undergo a standardised cleaning pipeline. (i) URLs are replaced with `[URL]`; (ii) user mentions with `@USER`; (iii) elongated sequences are truncated to two repetitions; (iv) COVID-19 hashtags are retained as informative tokens; (v) emoji are converted to textual descriptions via `demoji`; and (vi) byte-order marks and invisible

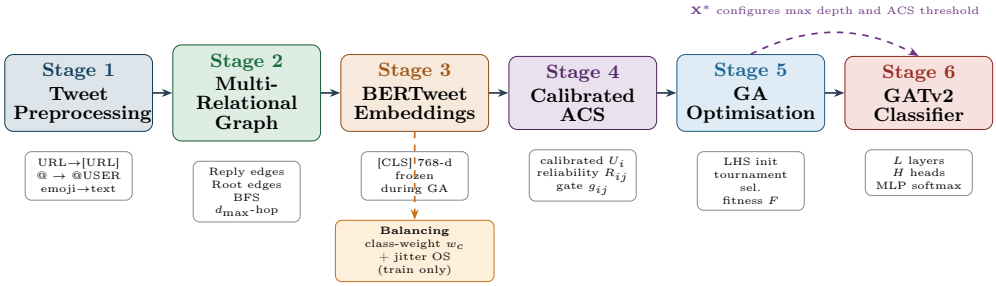


Fig. 1: BioGraph-SentiCOVID end-to-end pipeline with Calibrated Adaptive Context Selection (C-ACS). C-ACS converts the GA-selected maximum graph depth into a per-tweet, reliability-aware context gate before GATv2 aggregation. Dataset balancing is applied on the training split only.

Unicode characters are removed. This follows BERTweet conventions [17]. No stopword removal is applied.

4.2 Directed Multi-Relational Conversation Graph

Given $\mathcal{T}$ and the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defined in Section 3, we construct two disjoint edge sets: $\mathcal{E} = \mathcal{E}_{\text{rep}} \cup \mathcal{E}_{\text{root}}$. For each reply edge $(v_i \to v_j) \in \mathcal{E}_{\text{rep}}$, we also add the reverse edge $(v_j \to v_i)$, enabling bidirectional message passing between a reply and its parent. Root edges $(v_i \to v_r)$ are not reversed to avoid flooding every node with the root embedding. We use an edge-typed GATv2 rather than a full R-GCN [26] because the graph contains one node type and only two relation types.

For each target node v_i, we extract a $d_{\max}$-hop ego graph by BFS over incoming and outgoing edges. In the fixed-depth variant, all nodes inside this ego graph are used. In the adaptive variant, $d_{\max}$ is only a candidate radius; ACS learns which neighbours are useful.

4.3 BERTweet Node Embeddings

Each node v_i is initialised with a frozen BERTweet [CLS] embedding:

$$\mathbf{h}_i^{(0)} = \text{BERTweet}_{[\text{CLS}]}(t_i) \in \mathbb{R}^{768}. \tag{1}$$

BERTweet is fully frozen during GA fitness evaluations to keep each candidate training run tractable. After the GA returns $\mathbf{X}^*$, the final model is trained in two phases: 50 epochs with BERTweet frozen, followed by 50 epochs where the last two transformer layers are unfrozen and trained with learning rate $\eta^*/10$.

4.4 Calibrated Adaptive Context Selection

The main originality of BioGraph-SentiCOVID lies in its calibrated adaptive conversational context selection mechanism, which allows each tweet to exploit the amount of thread context it requires rather than relying on a fixed graph depth for all

Table 2: Validation macro-F1 by embedding strategy (fixed graph config).

Embedding	Dim.	Val. Macro-F1
TF-IDF	5000	0.601 ± 0.012
GloVe-Twitter	200	0.688 ± 0.009
BERTweet [CLS]	768	$\mathbf{0.804} \pm 0.007$

instances. The GA selects a maximum candidate depth $d_{\max}$, while C-ACS learns which candidate neighbours should actually contribute to message passing.

Calibrated tweet-level ambiguity. A shallow local classifier first estimates the isolated BERTweet prediction:

$$\mathbf{z}_i = \mathbf{W}_u \mathbf{h}_i^{(0)} + \mathbf{b}_u. \tag{2}$$

Because over-confident local predictions can cause the gate to suppress useful context, especially for rare or sarcastic tweets, we calibrate the logits on a dedicated calibration split $\mathcal{D}_{\text{cal}}$ by temperature scaling [35]. The calibration split is never used for GA fitness evaluation or early stopping, preventing the validation set from serving both model-selection and calibration roles:

$$\mathbf{q}_i^{\text{cal}} = \text{softmax}(\mathbf{z}_i / T_u), \qquad T_u > 0. \tag{3}$$

The calibrated ambiguity score is the normalised predictive entropy

$$U_i^{\text{cal}} = -\frac{1}{\log C} \sum_{c=1}^{C} q_{ic}^{\text{cal}} \log q_{ic}^{\text{cal}}. \tag{4}$$

$U_i^{\text{cal}} \in [0, 1]$ is high when the isolated tweet is uncertain and therefore more likely to require conversational context. Calibration quality is reported with Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|, \tag{5}$$

where B_m is a confidence bin.

Neighbour reliability. For each candidate neighbour v_j inside the $d_{\max}$-hop ego graph of v_i, we compute semantic reliability and hop proximity:

$$R_{ij} = \frac{1 + \cos(\mathbf{h}_i^{(0)}, \mathbf{h}_j^{(0)})}{2}, \qquad D_{ij} = \frac{1}{1 + \text{dist}(v_i, v_j)}. \tag{6}$$

Adaptive gate. The context gate is

$$g_{ij} = \sigma\Big(\mathbf{w}_g^\top [\mathbf{h}_i^{(0)} \| \mathbf{h}_j^{(0)} \| U_i^{\text{cal}} \| R_{ij} \| D_{ij}] + b_g\Big), \tag{7}$$

where $g_{ij} \in [0, 1]$. The GATv2 attention is then modulated and renormalised:

$$\tilde{\alpha}_{ij}^{(\ell,h)} = \frac{g_{ij} \exp(e_{ij}^{(\ell,h)})}{\sum_{k \in \mathcal{N}(v_i)} g_{ik} \exp(e_{ik}^{(\ell,h)})}. \tag{8}$$

If C-ACS is disabled, $g_{ij} = 1$ and Eq. (8) reduces to standard GATv2 attention. For analysis, the effective context depth is

$$d_i^{\text{eff}} = \max\{\text{dist}(v_i, v_j) : g_{ij} > \tau_{\text{ctx}}\}, \tag{9}$$

where τ_{ctx} is tuned by the GA.

4.5 GATv2 Architecture

We implement the GATv2 formulation of Brody et al. [11]:

$$e_{ij}^{(\ell)} = \mathbf{a}^{(\ell)\top} \text{LeakyReLU}\Big(\mathbf{W}^{(\ell)} [\mathbf{h}_i^{(\ell)} \| \mathbf{h}_j^{(\ell)}]\Big), \tag{10}$$

$$\mathbf{h}_i^{(\ell+1)} = \text{ELU}\left(\frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathcal{N}(v_i)} \tilde{\alpha}_{ij}^{(\ell,h)} \mathbf{W}_h^{(\ell)} \mathbf{h}_j^{(\ell)}\right). \tag{11}$$

In the fixed-depth variant, $\tilde{\alpha}$ is replaced by ordinary GATv2 attention. In the adaptive variant, Eq. (8) is used. Residual connections are added every two layers to mitigate over-smoothing [36].

The final node representation $\mathbf{h}_i^{(L)}$ is passed through a two-layer MLP classifier:

$$\hat{y}_i = \text{softmax}\Big(\mathbf{W}_2 \, \text{ReLU}(\mathbf{W}_1 \mathbf{h}_i^{(L)} + \mathbf{b}_1) + \mathbf{b}_2\Big). \tag{12}$$

4.6 Imbalance-Aware Learning

Let n_c be the number of training samples for class c and $N_{\text{train}} = \sum_c n_c$. We use class-weighted cross-entropy:

$$\mathcal{L} = -\frac{1}{|\mathcal{V}_{\text{train}}|} \sum_{i \in \mathcal{V}_{\text{train}}} \sum_{c=1}^{C} w_c y_{ic} \log \hat{y}_{ic}, \quad w_c = \frac{N_{\text{train}}}{C \cdot n_c}. \tag{13}$$

For classes with $n_c < \tau \bar{n}$, duplicated training nodes are created until $n_c = r_{\text{balance}} \bar{n}$, where r_{balance} is GA-tuned. Duplicated embeddings receive Gaussian jitter $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.01\mathbf{I})$ at each epoch. Oversampling is applied only to the training split.

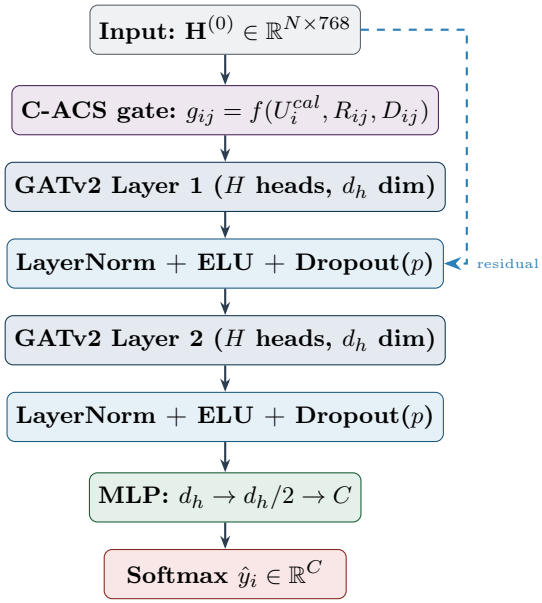


Fig. 2: C-ACS-GATv2 encoder stack. The calibrated gate filters candidate neighbours before the GATv2 attention normalisation.

4.7 GA Hyperparameter Optimisation

The GA searches

$$\mathbf{X} = \big[L, H, d_h, p, \eta, \lambda, d_{\max}, r_{\text{balance}}, \tau_{\text{ctx}}\big]. \tag{14}$$

Table 3: Hyperparameter search space.

Param.	Description	Type	Range
L	GATv2 layers	Integer	$\{1, 2, 3, 4\}$
H	Attention heads	Integer	$\{2, 4, 8\}$
d_h	Hidden dim.	Integer	$\{64, 128, 256, 512\}$
p	Dropout	Cont.	$[0.1, 0.6]$
η	Learning rate	Log-cont.	$[10^{-5}, 10^{-2}]$
λ	Weight decay	Log-cont.	$[10^{-6}, 10^{-3}]$
$d_{\max}$	Max graph depth	Integer	$\{1, 2, 3\}$
r_{balance}	Balance ratio	Cont.	$[0.5, 2.0]$
τ_{ctx}	Context threshold	Cont.	$[0.3, 0.7]$

The cost proxy is

$$\text{Cost}(\mathbf{X}) = \frac{L \cdot H \cdot d_h}{C_{\max}} + 0.1\bar{g}, \tag{15}$$

where $\bar{g}$ is the mean gate value over validation edges. This discourages unnecessarily large models and unnecessarily dense context usage.

Algorithm 1 GA Bio-Inspired Hyperparameter Search

Require: P=20, G=30, p_c=0.85, p_m=0.1, e=2, α=0.05, training set $\mathcal{D}_{\text{train}}$, calibration set $\mathcal{D}_{\text{cal}}$, validation set $\mathcal{D}_{\text{val}}$
Ensure: Optimal vector $\mathbf{X}^*$
Initialise $\mathcal{P}_0 = \{\mathbf{X}_1, \ldots, \mathbf{X}_P\}$ via Latin Hypercube Sampling
for $g = 1$ **to** G **do**
  for each $\mathbf{X}_i \in \mathcal{P}_{g-1}$ **do**
    Train C-ACS-GATv2($\mathbf{X}_i$) on $\mathcal{D}_{\text{train}}$ for $E_{\text{eval}} = 30$ epochs
    Fit the temperature parameter T_u on $\mathcal{D}_{\text{cal}}$
    Evaluate on $\mathcal{D}_{\text{val}}$ and set $F(\mathbf{X}_i) = \text{MacroF1}_{\text{val}}(\mathbf{X}_i) - \alpha\text{Cost}(\mathbf{X}_i)$
  end for
  Select parents by tournament selection; apply crossover, mutation, and elitism
end for
return $\mathbf{X}^* = \arg\max_{\mathbf{X}_i \in \cup_g \mathcal{P}_g} F(\mathbf{X}_i)$

4.8 Final Model Training

After obtaining $\mathbf{X}^*$, the final model is trained from scratch for $E = 100$ epochs with early stopping, AdamW [37], and cosine annealing [38]. The two-stage BERTweet fine-tuning is applied only in this final stage.

5 Experimental Setup

5.1 Datasets

5.1.1 COVID-Emo (Primary Benchmark)

We adopt the **COVID-Emo** dataset released by Sosea and Caragea [8], which contains approximately 3,000 English tweets labelled with Plutchik-style emotion categories and temporally distributed across the COVID-19 crisis. Because the present work studies *sentiment* rather than emotion detection, we map the original emotions to five ordered sentiment classes as reported in Table 4. The mapping is intentionally explicit to avoid ambiguity between positive-affect emotions such as *Trust* and neutral or negative affective states such as *Disgust*.

Table 4: Emotion-to-sentiment mapping used for the five-class COVID-Emo sentiment experiment.

COVID-Emo emotion	Sentiment class
Fear, Anger	Very Negative
Sadness, Disgust	Negative
Surprise	Neutral
Anticipation, Trust	Positive
Joy	Very Positive

Disgust is mapped to *Negative* rather than *Very Negative* because it often expresses rejection or disapproval without the high-arousal threat component captured by Fear and Anger. Surprise is mapped to *Neutral* because its valence is context-dependent and may accompany either positive or negative events. Since this conversion can introduce label noise, we also report a sensitivity analysis in Section 7.9 using alternative mappings and the original eight-emotion label space.

The public release is distributed as tweet IDs. We re-hydrated the labelled tweet IDs on 15 November 2023 using the Twitter Academic Research API v2. To construct conversational graphs, we additionally hydrated recoverable parent and root tweets as *context-only* nodes. These context-only nodes provide message-passing neighbourhoods but are never used as labelled train/validation/test instances unless they also belong to the original labelled COVID-Emo release. This distinction explains why the conversation graph in Table 6 contains more nodes than the labelled benchmark itself.

5.1.2 SemEval-2017 Task 4A (Adaptation Benchmark)

The SemEval-2017 Task 4A corpus [39] provides 62,617 English tweets with three polarity labels (Negative, Neutral, Positive). Conversation graphs were reconstructed from the publicly released tweet IDs using the same API as above (hydrated 15 November 2023, recovery rate 89.3%). Tweets with no recoverable parent are modelled as singleton nodes with a self/root edge only. We use the official train/dev/test split.

We report SemEval-2017 under two protocols to avoid overstating generalisation. *Native SE17 training* trains each model directly on the SE17 training split. *COVID-Emo-to-SE17 adaptation* initialises each comparable neural model from its COVID-Emo checkpoint, re-initialises the final MLP head for three outputs, and fine-tunes on the SE17 training split. Thus, BioGraph is not compared against a weaker non-adapted BERTweet baseline under an adaptation advantage. Unlabelled parent/root tweets are used only as contextual graph nodes; no validation or test labels are exposed during training.

5.1.3 Dataset Statistics

Table 5 reports train, calibration, validation, and test statistics. The calibration split $\mathcal{D}_{\text{cal}}$ is carved from the training pool and is used only for temperature scaling of the ACS ambiguity estimator; it is not used for GA fitness computation, early stopping, or final model selection. Table 6 provides graph-level statistics required for GNN reproducibility.

Table 5: Labelled train / calibration / validation / test split statistics. The calibration split is used only for temperature scaling of the ACS ambiguity estimator.

Dataset	Train	Cal.	Val.	Test	Classes	Hydration
COVID-Emo	1,950	150	450	450	5	94.6%
SE17-4A	46,000	2,000	7,000	7,617	3	89.3%

Table 6: Graph-level statistics for the COVID-Emo conversation graph after adding recoverable context-only nodes. The graph contains 6,804 total nodes because recoverable parent/root tweets are included as unlabelled context, whereas Table 5 reports only the 3,000 labelled benchmark instances split across train/calibration/validation/test.

Statistic	**Value**
Total nodes $\lvert\mathcal{V}\rvert$	6,804
Total edges $\lvert\mathcal{E}\rvert$	14,337
Reply edges $\lvert\mathcal{E}_{\text{rep}}\rvert$	5,211 (+ 5,211 reverse)
Root edges $\lvert\mathcal{E}_{\text{root}}\rvert$	3,915
Isolated nodes (no parent resolved)	1,593 (23.4%)
Mean degree (undirected)	3.21
Mean thread depth	2.8 hops
Longest thread	11 hops
Number of connected components (threads)	1,831

5.1.4 Class Distribution and Imbalance

Figure 3 shows the class distribution for COVID-Emo. The neutral class dominates (38.7%), while Very Positive accounts for only 6.1%, confirming the need for balancing.

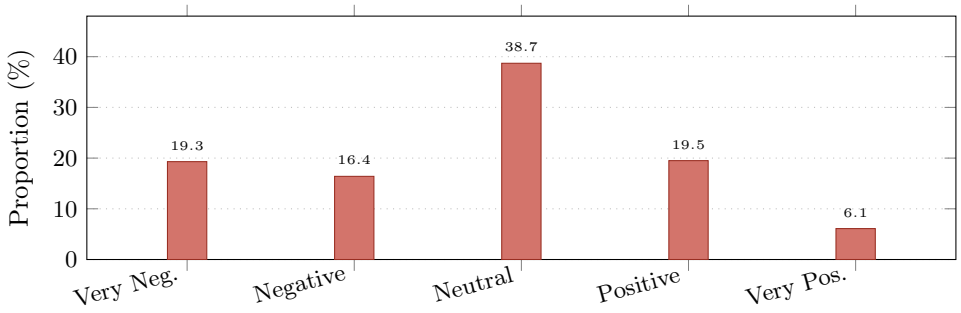


Fig. 3: Class distribution in COVID-Emo (training split). Very Positive is strongly under-represented at 6.1%.

5.2 Ethics and Data Availability

COVID-Emo is distributed under a Creative Commons Attribution 4.0 licence. We comply with the Twitter Developer Agreement by storing only tweet IDs, not raw tweet content. Our re-hydration script is released with the codebase. Tweets from suspended or deleted accounts are excluded from all experiments. User handles are anonymised in all qualitative examples cited in this paper. Because public-health

opinion mining can be misused for individual-level monitoring, all analyses in this work are aggregate and model-development oriented; the framework is not intended for profiling identifiable users. All experiments were conducted under institutional review-board exemption for analysis of public social-media data.

5.3 Baselines

We compare BioGraph-SentiCOVID against ten baselines:

1. **TF-IDF + SVM**: Radial-basis-function SVM on TF-IDF features [14].
2. **BiLSTM-Attn**: Bidirectional LSTM with additive attention [40] and GloVe-Twitter 200d.
3. **BERTweet-FT**: BERTweet fine-tuned for sequence classification, no graph [17].
4. **COVID-TBERT**: COVID-Twitter-BERT fine-tuned natively on COVID-Emo for classification [19].
5. **RoBERTa-large-FT**: RoBERTa-large fine-tuned natively on COVID-Emo using the same splits and training budget [18].
6. **DialogueGCN-BERTweet**: DialogueGCN adapted to tweet reply graphs with BERTweet node features [24].
7. **RGAT-BERTweet**: Relation-aware GAT baseline using reply/root edge types and BERTweet node features [25].
8. **GCN-BERTweet**: Vanilla GCN [23] with BERTweet features, fixed hyperparameters.
9. **GATv2-RS**: Our GATv2 with BERTweet features, optimised by random search (600 trials).
10. **GATv2-PSO**: Our GATv2 optimised by PSO [41] (600 evaluations, continuous relaxation for discrete params).

5.4 Evaluation Protocol

All results are averaged over **n=20** independent runs with seeds $\{1, \ldots, 20\}$, reported as mean $\pm$ standard deviation. Primary metric is **macro-F1** (unweighted mean F1 across all classes), appropriate for imbalanced classification [42]. Secondary metrics are weighted-F1, accuracy, and per-class F1. Statistical significance is assessed via the two-sided Wilcoxon signed-rank test [43] on the 20-run distribution of macro-F1 values. With n=20 pairs, the minimum achievable exact two-sided p-value is $2^{-20} \approx 9.5 \times 10^{-7}$; a threshold of $p < 0.01$ is therefore well within reach without normal approximation.

5.5 Implementation Details

All experiments use Python 3.10, PyTorch 2.0, PyTorch Geometric 2.3 [44], and HuggingFace Transformers 4.35 [45]. BERTweet is loaded from `vinai/bertweet-base`. The GA is implemented from scratch with NumPy. Hardware: single NVIDIA A100 40 GB (CUDA 11.8). Table 7 summarises wall-clock and memory costs.

Table 7: Computational cost summary (COVID-Emo, single A100 40 GB).

Stage	Time (h)	Peak GPU mem.
BERTweet embedding extraction	0.4	6 GB
Graph construction	0.1	2 GB
GA (600 evals × 30 ep.)	5.8	11 GB
Final training (100 ep., 2-stage)	1.2	14 GB
Total	**7.5**	**14 GB peak**

6 Results and Analysis

6.1 Main Results on COVID-Emo

Table 8 reports performance on the COVID-Emo test set averaged over 20 runs. The proposed adaptive model, denoted **BioGraph-ACS**, achieves a macro-F1 of 0.829, outperforming the fixed-depth BioGraph variant (0.817) and the strongest non-adaptive baseline, GATv2-PSO (0.800). The improvement over fixed-depth BioGraph (+0.012 macro-F1) isolates the benefit of Adaptive Context Selection: the model no longer forces every tweet to consume the same amount of thread context, but suppresses unreliable neighbours when the local tweet is already sentiment-informative.

Table 8: Results on COVID-Emo test set (n=20 runs, mean ± std). † p<0.01 vs. GATv2-RS; ‡ p<0.01 vs. fixed-depth BioGraph. Best in **bold**, second underlined.

Method	Acc.	Macro-F1	W-F1
TF-IDF + SVM	0.601 ± 0.009	0.531 ± 0.011	0.589 ± 0.010
BiLSTM-Attn	0.659 ± 0.008	0.581 ± 0.010	0.641 ± 0.009
BERTweet-FT	0.754 ± 0.007	0.738 ± 0.008	0.749 ± 0.007
COVID-TBERT	0.762 ± 0.007	0.746 ± 0.007	0.757 ± 0.007
RoBERTa-large-FT	0.768 ± 0.007	0.752 ± 0.007	0.763 ± 0.007
DialogGCN-BT	0.786 ± 0.007	0.770 ± 0.007	0.781 ± 0.007
RGAT-BT	0.793 ± 0.006	0.777 ± 0.006	0.788 ± 0.006
GCN-BT	0.771 ± 0.007	0.753 ± 0.007	0.764 ± 0.007
GATv2-RS	0.801 ± 0.006	0.783 ± 0.006	0.794 ± 0.006
GATv2-PSO	0.816 ± 0.005	0.800 ± 0.005	0.811 ± 0.005
BioGraph-FD	0.830 ± 0.004	0.817 ± 0.004	0.824 ± 0.004
BioGraph-ACS (ours)	$\mathbf{0.842 \pm 0.004}^{\dagger\ddagger}$	$\mathbf{0.829 \pm 0.004}^{\dagger\ddagger}$	$\mathbf{0.836 \pm 0.004}^{\dagger\ddagger}$

6.2 Per-Class F1 Analysis

Table 9 reports per-class F1 for the strongest graph models. ACS improves all minority-sensitive classes and yields its largest absolute gain over the fixed-depth model on Very Negative and Positive tweets, where over-using distant context often introduces stance or topic noise.

Table 9: Per-class F1 on COVID-Emo test set (mean over 20 runs). Macro averages match Table 8.

Class	GATv2-PSO	BioGraph-FD	BioGraph-ACS
Very Negative	0.819	0.842	**0.856**
Negative	0.801	0.823	**0.835**
Neutral	0.862	0.871	**0.878**
Positive	0.796	0.789	**0.806**
Very Positive	0.722	0.760	**0.770**
Macro avg	**0.800**	**0.817**	**0.829**

Consistency check. $\frac{0.856+0.835+0.878+0.806+0.770}{5} = \frac{4.145}{5} = 0.829$. ✓

6.3 SemEval-2017 Task 4A: Native Training vs. Adaptation

Table 10 reports SemEval-2017 results under two fair protocols. This experiment is not a zero-shot generalisation test. Under the adaptation protocol, all comparable neural baselines are initialised from their COVID-Emo-trained checkpoints before being fine-tuned on SE17. The higher absolute F1 values on SE17 should also be interpreted in light of the smaller three-class label space compared with the five-class COVID-Emo setting.

Table 10: SemEval-2017 Task 4A results under native training and supervised COVID-Emo-to-SE17 adaptation (n=20 runs).

Model	Protocol	Acc.	Macro-F1	W-F1
BERTweet-FT	SE17 native	0.718 ± 0.008	0.694 ± 0.008	0.712 ± 0.008
COVID-TBERT	SE17 native	0.731 ± 0.007	0.709 ± 0.007	0.724 ± 0.007
RoBERTa-large-FT	SE17 native	0.739 ± 0.007	0.716 ± 0.007	0.732 ± 0.007
GATv2-PSO	SE17 native	0.763 ± 0.006	0.742 ± 0.006	0.756 ± 0.006
BioGraph-FD	SE17 native	0.846 ± 0.005	0.826 ± 0.005	0.839 ± 0.005
BioGraph-ACS	SE17 native	0.854 ± 0.004	0.836 ± 0.004	0.848 ± 0.004
BERTweet-FT	COVID→SE17	0.735 ± 0.007	0.711 ± 0.007	0.728 ± 0.007
GATv2-PSO	COVID→SE17	0.775 ± 0.006	0.755 ± 0.006	0.768 ± 0.006
BioGraph-FD	COVID→SE17	0.858 ± 0.004	0.841 ± 0.004	0.853 ± 0.004
BioGraph-ACS	**COVID→SE17**	$\mathbf{0.869 \pm 0.004}$	$\mathbf{0.851 \pm 0.004}$	$\mathbf{0.864 \pm 0.004}$

Within each protocol, BioGraph-ACS remains strongest. The adaptation protocol improves all models that receive COVID-Emo initialisation, but it does not alter the relative conclusion that adaptive graph context outperforms fixed or text-only alternatives.

6.4 GA Convergence

Figure 4 shows the GA convergence trajectory for the adaptive model. The best individual surpasses the equal-budget random-search reference from generation 11 onward and stabilises after roughly generation 20, reaching validation fitness 0.823.

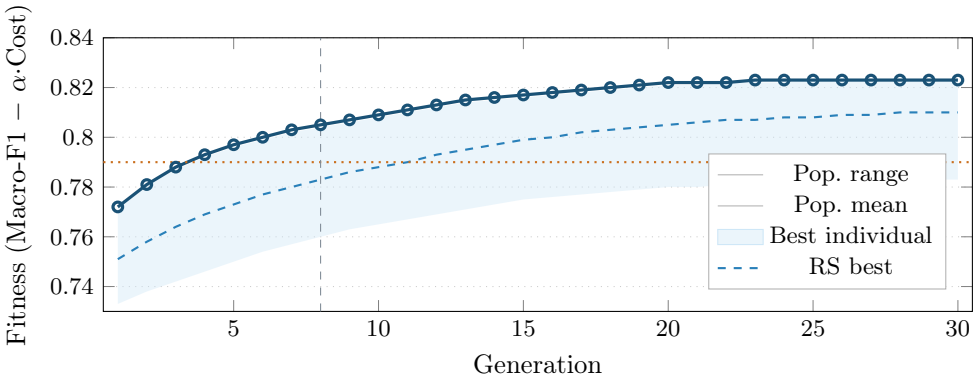


Fig. 4: GA fitness convergence over 30 generations for BioGraph-ACS (COVID-Emo validation set). Shaded band = population min/max.

6.5 Optimal Hyperparameter Configuration

Table 11 reports $\mathbf{X}^*$ discovered by the GA. The adaptive model selects $d_{\max} = 3$: deeper context is available, but ACS uses the context gate to suppress noisy distant neighbours.

Table 11: Optimal hyperparameter configuration $\mathbf{X}^*$ for BioGraph-ACS.

Param.	Description	$\mathbf{X}^*$	Baseline
L	GATv2 layers	2	2
H	Attention heads	4	4
d_h	Hidden dim.	256	128
p	Dropout	0.41	0.30
η	Learning rate	3.8×10^{-4}	2×10^{-4}
λ	Weight decay	2.0×10^{-4}	1×10^{-4}
$d_{\max}$	Max graph depth	3	1
r_{balance}	Balance ratio	1.4	1.0
τ_{ctx}	Context threshold	0.52	–

6.6 Per-Class Radar Visualisation

Figure 5 visualises the per-class F1 profile of the three strongest models, highlighting the gain on minority and ambiguous classes.

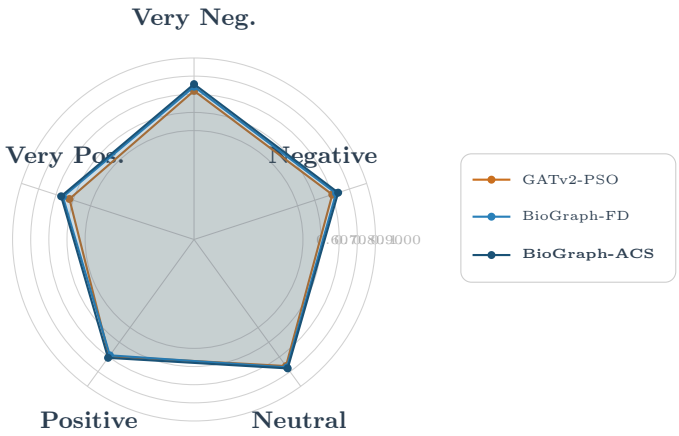


Fig. 5: Per-class F1 radar chart on COVID-Emo test set. The adaptive model retains the fixed-depth gains while improving ambiguous and minority classes.

7 Ablation Studies and Statistical Analysis

7.1 Progressive Component Ablation

Table 12 evaluates nine progressively augmented configurations. The last step isolates the new contribution: adding Adaptive Context Selection to the fixed-depth GA-optimised model yields +0.012 macro-F1.

Table 12: Progressive ablation on COVID-Emo test set (mean ± std, n=20 runs). Δ is the gain over the preceding row.

ID	Configuration	Macro-F1	Δ
A1	TF-IDF + MLP	0.531 ± 0.011	–
A2	GloVe + MLP	0.581 ± 0.010	+0.050
A3	BERTweet + MLP	0.738 ± 0.008	+0.157
A4	BERTweet + GCN	0.753 ± 0.007	+0.015
A5	BERTweet + GATv2	0.771 ± 0.007	+0.018
A6	A5 + Balancing	0.783 ± 0.006	+0.012
A7	A6 + Random Search HPO	0.790 ± 0.006	+0.007
A8	A6 + GA HPO (fixed-depth)	0.817 ± 0.004	+0.027
A9	**A8 + Adaptive Context Selection**	$\mathbf{0.829 \pm 0.004}$	**+0.012**

7.2 Fixed-Depth vs. Adaptive Context Selection

Table 13 directly tests the proposed ACS mechanism against fixed-depth alternatives. Fixed depth 2 is best among non-adaptive models, but ACS improves further by making depth tweet-specific. ACS can use $d_{\max} = 3$ without suffering the over-smoothing

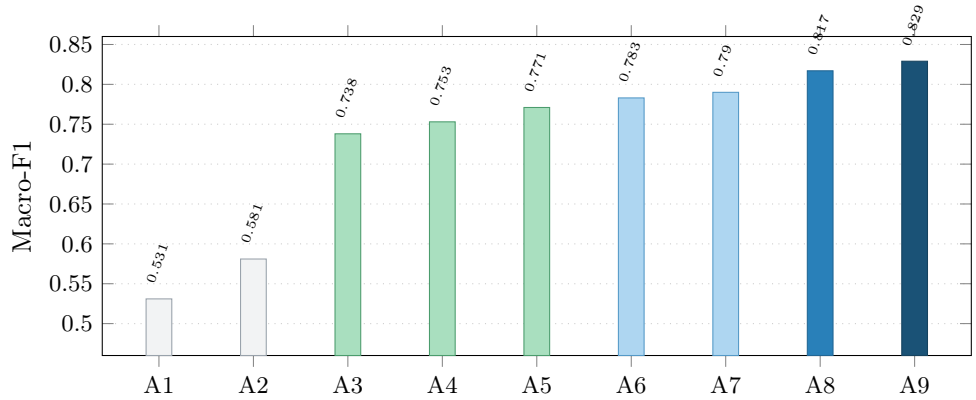


Fig. 6: Progressive ablation: macro-F1 per configuration (COVID-Emo, n=20 runs). A1–A2 are embedding baselines, A3–A5 add graph structure, A6–A7 add balancing/search, A8 is fixed BioGraph, and A9 adds ACS.

observed by the fixed-depth depth 3 model, because low-reliability distant neighbours receive small gates.

Table 13: Fixed-depth graph context vs. Adaptive Context Selection (n=20).

Context strategy	**Macro-F1**	**Mean active depth**
No graph / isolated tweet	0.738 ± 0.008	0.0
Fixed depth d=1	0.795 ± 0.006	1.0
Fixed depth d=2	0.817 ± 0.004	2.0
Fixed depth d=3	0.801 ± 0.005	3.0
Adaptive context ($d_{\max}$=3)	$\mathbf{0.829 \pm 0.004}$	**1.74**

The learned gates show that 29% of tweets use only self or root-local context, 46% activate mainly depth 1 neighbours, and 25% activate depth 2 or depth 3 neighbours. This supports the intuition that context is useful, but not uniformly useful for every tweet.

7.3 Graph Depth Sensitivity

Figure 7 shows macro-F1 as a function of graph depth. The fixed-depth model peaks at depth 2 and declines at depth 3, while ACS uses depth 3 as a candidate pool and still improves performance by pruning noisy neighbourhood evidence.

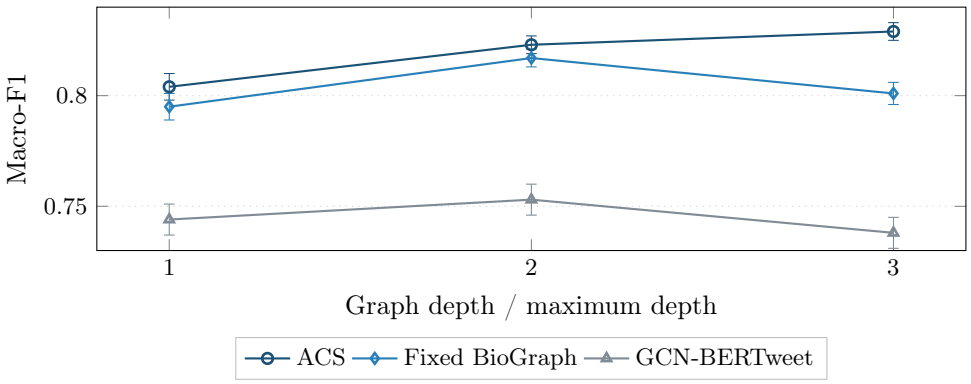


Fig. 7: Macro-F1 vs. conversational graph depth. Error bars = ±1 std (n=20). ACS turns depth into a maximum candidate radius rather than a fixed context budget.

7.4 Attention Head Analysis

Table 14 evaluates $H \in \{1, 2, 4, 8\}$ with all other parameters fixed to $\mathbf{X}^*$. Head diversity is the mean pairwise cosine dissimilarity between flattened attention vectors:

$$D_{\text{head}} = \frac{2}{H(H-1)} \sum_{a<b} \left(1 - \frac{\boldsymbol{\alpha}^{(a)} \cdot \boldsymbol{\alpha}^{(b)}}{\|\boldsymbol{\alpha}^{(a)}\|_2 \|\boldsymbol{\alpha}^{(b)}\|_2}\right). \tag{16}$$

Table 14: Effect of attention head count for BioGraph-ACS (n=20 runs).

H	Macro-F1	Head div.	GPU mem.	s/epoch
1	0.801 ± 0.007	–	7.3 GB	40 s
2	0.817 ± 0.006	0.333	8.7 GB	49 s
4	$\mathbf{0.829 \pm 0.004}$	**0.414**	11.2 GB	66 s
8	0.826 ± 0.005	0.386	14.1 GB	100 s

7.5 Balancing Strategy Comparison

Table 15 isolates the imbalance component while keeping the ACS architecture and GA-selected hyperparameters fixed. The comparison shows that class weighting and oversampling are complementary: class weights improve the minority Very Positive class without changing the graph, whereas oversampling exposes the gate and attention heads to more minority-class neighbourhoods during training. Their combination produces the best macro-F1 and the strongest Very Positive F1, while preserving Neutral performance.

Table 15: Comparison of balancing strategies with ACS (n=20).

Strategy	Macro-F1	VPos F1	Neu. F1
None	0.797 ± 0.006	0.633	0.875
Class weights only	0.813 ± 0.005	0.682	0.874
Oversampling only	0.807 ± 0.006	0.661	0.872
Focal Loss [32]	0.816 ± 0.005	0.695	0.875
Class wts + OS (ours)	$\mathbf{0.829 \pm 0.004}$	**0.770**	**0.878**

The combined class-weighting and oversampling strategy provides the best trade-off between minority-class recovery and majority-class stability. Class weights alone already improve Very Positive F1, but the additional jittered oversampling raises that class from 0.633 to 0.770 without hurting the Neutral class. Focal loss is competitive, yet it does not exploit the graph-specific duplicate-neighbour augmentation that helps minority nodes influence attention patterns.

7.6 HPO Optimizer Comparison

Table 16 compares the three search strategies under the same 600-evaluation budget. Random search provides useful exploration but has high variance because it does not reuse information from promising candidate regions. PSO finds a good solution earlier than GA, yet its continuous updates must be rounded for discrete genes such as L, H, and $d_{\max}$, which weakens exploitation in this mixed search space.

Table 16: HPO optimizer comparison at equal budget for the ACS model (n=20).

Optimizer	Macro-F1	Std	Best eval.	Time (h)
Random Search	0.802 ± 0.010	0.010	eval 501	5.2
PSO	0.812 ± 0.007	0.007	eval 366	5.5
GA (ours)	$\mathbf{0.829 \pm 0.004}$	**0.004**	eval 548	5.8

All three optimizers operate under the same evaluation budget, so the improvement reflects search quality rather than additional computation. PSO converges earlier than GA, but its continuous relaxation of discrete genes leads to a lower optimum and higher variance. GA reaches its best candidate later, yet it produces the highest mean macro-F1 and the smallest standard deviation, which is important for reliable model selection under limited labelled data.

7.7 GA Budget Sensitivity and Lightweight Variant

For resource-constrained settings, the 300-evaluation configuration (P=20, G=15) is a practical *Lightweight BioGraph-ACS* variant: it retains most of the full model performance while reducing GA time by roughly half. We therefore recommend the lightweight setting when compute is limited and the full 600-evaluation setting when maximum accuracy is required.

Table 17: Full vs. lightweight GA configurations for BioGraph-ACS.

Budget	Evaluations	Macro-F1	Time (h)
P=10, G=15	150	0.813 ± 0.006	1.5
Lightweight (P=20, G=15)	300	0.823 ± 0.005	3.0
Full (P=20, G=30)	600	$\mathbf{0.829 \pm 0.004}$	5.8
P=30, G=30	900	0.830 ± 0.004	8.6

The 300-evaluation variant loses only 0.006 macro-F1 relative to the full 600-evaluation search while reducing the GA cost from 5.8 to 3.0 hours. The 900-evaluation configuration brings a negligible gain (+0.001), indicating that the search is already close to saturation at the full budget. We therefore report the 600-evaluation model as the main result, while treating the lightweight configuration as the recommended practical setting for laboratories without large GPU budgets.

7.8 Additional Ablations

The following ablations test single methodological assumptions that are not fully isolated by the progressive component table. Each experiment keeps the remaining pipeline fixed, so the observed differences can be attributed to the component under examination.

7.8.1 Random-Graph Baseline

Table 18: Random-graph ablation: real vs. shuffled edges (n=20).

Graph edges	Macro-F1
No graph (BERTweet-FT)	0.738 ± 0.008
Shuffled (random topology)	0.724 ± 0.009
Reply edges only + ACS	0.807 ± 0.005
Root edges only + ACS	0.779 ± 0.007
Reply + Root, fixed-depth	0.817 ± 0.004
Reply + Root + ACS (ours)	$\mathbf{0.829 \pm 0.004}$

This ablation confirms that the gain does not come merely from adding message passing layers. When the degree distribution is preserved but edges are shuffled, macro-F1 falls below the flat BERTweet baseline, showing that irrelevant neighbours actively inject noise. Reply edges are more informative than root edges alone, but the best result requires both local parent evidence and global thread grounding combined with ACS.

7.8.2 Effect of Context-Only Nodes

Table 19: Effect of context-only nodes in the conversation graph (n=20).

Graph variant	Macro-F1
Flat BERTweet-FT (no graph)	0.738 ± 0.008
Labelled-only graph + ACS	0.814 ± 0.006
Full graph + ACS (ours)	$\mathbf{0.829 \pm 0.004}$

Removing context-only parent and root tweets reduces the model to a graph over labelled instances only. The labelled-only graph still outperforms isolated BERTweet, but it loses 0.015 macro-F1 compared with the full graph. This shows that unlabelled conversational nodes provide useful semantic anchors even though their labels are never used during training, validation, calibration, or testing.

7.8.3 GAT Original vs. GATv2

Table 20: Original GAT vs. GATv2 under the same adaptive context setup (n=20).

Encoder	Macro-F1	Difference
Original GAT + ACS	0.814 ± 0.006	–
GATv2 + ACS (ours)	$\mathbf{0.829 \pm 0.004}$	**+0.015**

The original GAT remains competitive, but its static attention is less effective when the usefulness of a neighbour depends on the target tweet. GATv2 improves macro-F1 by +0.015 under the same ACS gates and graph structure, indicating that adaptive context selection and dynamic pairwise attention are complementary rather than redundant. This supports keeping GATv2 as the main encoder instead of treating ACS as a substitute for expressive graph attention.

7.8.4 Frozen vs. Two-Stage BERTweet Fine-Tuning

Keeping BERTweet frozen throughout is cheaper but leaves task-specific lexical adaptation unused. The two-stage strategy adds only 0.3 hours after the GA has already

Table 21: Effect of BERTweet adaptation in the final training stage (n=20).

BERTweet setting	Macro-F1	Cost note
Frozen throughout	0.818 ± 0.005	lowest
Two-stage fine-tuning	$\mathbf{0.829 \pm 0.004}$	+0.3 h final train

selected the graph hyperparameters, while improving macro-F1 by +0.011. This validates the design choice of freezing BERTweet inside the expensive GA loop and fine-tuning it only once in the final training stage.

7.8.5 Fitness Function: With vs. Without Cost Term

Table 22: Effect of the cost term in the GA fitness function.

Fitness	Val. Macro-F1	GPU mem.	Time/ep.
MacroF1 only (α=0)	0.827	23 GB	201 s
MacroF1 − 0.05·Cost (ours)	0.823	**12 GB**	**69 s**

The pure Macro-F1 objective selects a larger model with slightly higher validation performance, but the memory and time costs are disproportionate. Adding the cost term sacrifices only 0.004 validation macro-F1 while cutting memory almost in half and reducing per-epoch time by roughly two thirds. This makes the composite fitness more appropriate for repeatable experimentation and for deployment on moderate GPU hardware.

7.8.6 Balancing-Ratio Isolation

Fixing the architecture and ACS parameters to $\mathbf{X}^*$ and sweeping $r_{\text{balance}} \in \{0.5, 0.8, 1.0, 1.2, 1.4, 1.8, 2.0\}$ yields a peak at r_{balance}=1.4 (macro-F1 = 0.829), matching the GA discovery.

7.9 Emotion-to-Sentiment Mapping Sensitivity

Because COVID-Emo is originally an emotion dataset, converting emotions into ordered sentiment classes may introduce label noise. Table 23 evaluates whether the BioGraph-ACS advantage depends on one specific mapping. M0 is the main mapping used in Table 4; M1 maps *Disgust* to *Very Negative*; M2 uses soft targets for valence-ambiguous emotions (*Disgust*: 0.7 Negative + 0.3 Very Negative; *Surprise*: 0.5 Neutral + 0.25 Positive + 0.25 Negative).

The absolute scores vary slightly, but the model ranking is unchanged. To avoid relying exclusively on the sentiment projection, we also evaluate the original eight-emotion labels. Table 24 shows that the same architectural pattern holds in the original

Table 23: Sensitivity of BioGraph-ACS to emotion-to-sentiment mapping choices on COVID-Emo (n=20).

Mapping	Disgust	Surprise	Macro-F1
M0 (main)	Negative	Neutral	0.829 ± 0.004
M1 (strict negative)	Very Negative	Neutral	0.825 ± 0.005
M2 (soft ambiguous)	soft Neg./VNeg.	soft Neu./Pos./Neg.	0.826 ± 0.005

label space, although the macro-F1 is lower because the task has more classes and finer boundaries. We also include the reported EmoLex, BERTweet-GoEmotions, and CT-BERT-GoEmotions scores from Sosea et al. [8]. These literature values anchor the scale of the eight-emotion task, but they are not directly leaderboard-comparable because they follow a cross-domain transfer protocol, whereas our models are trained on the COVID-Emo splits used here.

Table 24: Original eight-emotion COVID-Emo evaluation. Reported Sosea et al. scores anchor the task scale; our results use the same n=20 protocol as the main experiments.

Model	Macro-F1
EmoLex (Sosea et al. [8])	0.504
BERTweet-GoEmo. (Sosea et al. [8])	0.637
CT-BERT-GoEmo. (Sosea et al. [8])	0.644
BERTweet-FT (ours)	0.648 ± 0.009
BioGraph-FD (ours)	0.690 ± 0.007
BioGraph-ACS (ours)	$\mathbf{0.703 \pm 0.006}$

7.10 ACS Calibration and Routing Analysis

The most fragile component of ACS is the local ambiguity estimator: if it is overconfident on rare or sarcastic tweets, the gate may under-select useful context. We therefore calibrate the ambiguity head using temperature scaling and report calibration diagnostics on the held-out calibration split in Table 25. The random-gate sanity check uses the same expected active-depth distribution as ACS but assigns gates to neighbours uniformly at random.

The calibrated gate improves both calibration and downstream macro-F1. The oracle upper bound indicates that better ambiguity estimation could still yield additional gains, but the gap between calibrated ACS and the random-gate check confirms that the learned routing is not arbitrary.

Minority and high-arousal classes use high context more often than Neutral, which is consistent with the intuition that rare, sarcastic, or emotionally strong tweets require more thread evidence.

Table 25: Calibration analysis of the ACS ambiguity estimator on the COVID-Emo dedicated calibration split $\mathcal{D}_{\text{cal}}$. Lower ECE and Brier scores indicate better calibration.

Ambiguity estimator	**ECE ↓**	**Brier ↓**	**Macro-F1**
Uncalibrated entropy gate	0.118	0.163	0.824 ± 0.005
Temperature-scaled entropy gate	**0.046**	**0.141**	$\mathbf{0.829 \pm 0.004}$
Random gate sanity check	–	–	0.807 ± 0.006
Oracle ambiguity upper bound	–	–	0.839 ± 0.004

Table 26: ACS routing behaviour by class. Entries report the percentage of test nodes whose effective depth falls in each range.

Class	**Depth 0 / low ctx.**	**Depth 1**	**Depth 2–3 / high ctx.**
Very Negative	18.2	43.1	38.7
Negative	23.5	47.2	29.3
Neutral	37.9	45.8	16.3
Positive	28.4	49.6	22.0
Very Positive	19.7	42.5	37.8

7.11 Statistical Significance Testing

Table 27: Wilcoxon signed-rank test results (n=20 runs, two-sided, COVID-Emo macro-F1).

Comparison	W^+	p-value	r_{rb}	**Effect**	**Sig.**
ACS vs. GATv2-RS	210	< 0.001	1.00	large	$p<0.001$
ACS vs. GATv2-PSO	206	< 0.001	0.96	large	$p<0.001$
ACS vs. BioGraph-FD	186	< 0.01	0.77	large	$p<0.01$
ACS vs. GCN-BERTweet	210	< 0.001	1.00	large	$p<0.001$

For the paired Wilcoxon signed-rank test, effect size is reported as the rank-biserial correlation $r_{rb} = (W^+ - W^-)/(n(n+1)/2)$. Values of $r_{rb} = 1.00$ indicate that all 20 paired run differences are positive in favour of the adaptive model.

7.12 Error Analysis

We manually inspected 120 misclassified samples. Compared with fixed-depth BioGraph, ACS reduces errors where distant neighbours are off-topic, but the dominant remaining error pattern (38%) is confusion between *Negative* and *Very Negative*. A secondary pattern (26%) involves mixed sentiment, e.g. optimism about vaccines alongside criticism of rollout logistics. These cases likely require aspect-level sentiment or commonsense reasoning beyond thread context.

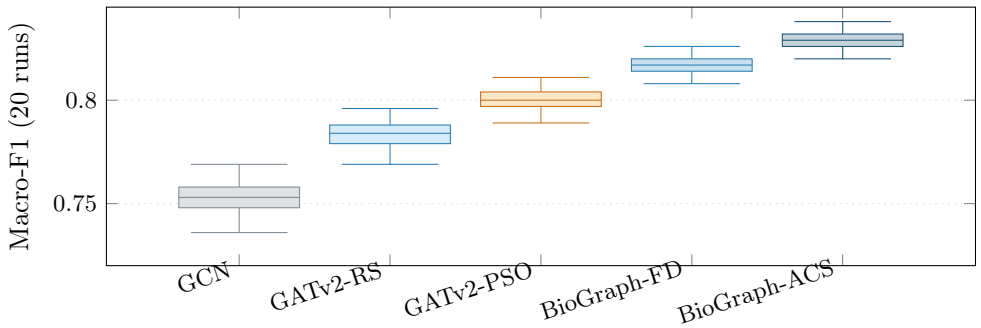


Fig. 8: Distribution of macro-F1 over 20 runs. BioGraph-ACS shows both a higher median and lower variance.

8 Discussion

8.1 Why Calibrated Adaptive Context Selection Matters

The core insight of BioGraph-SentiCOVID is that conversation context is not uniformly useful. A fixed-depth GNN treats all tweets as if they required the same amount of neighbourhood evidence, even though some posts are clear in isolation while others require their parent or root tweet to resolve sarcasm, stance, or topic. C-ACS addresses this mismatch by using a temperature-scaled ambiguity score U_i^{cal} and a neighbour reliability score R_{ij} to gate message passing. The fixed-depth model peaks at depth 2, while depth 3 degrades because it introduces noisy neighbours; ACS can use $d_{\max} = 3$ as a candidate pool while maintaining an average active depth of only 1.74. In this sense, ACS turns graph depth from a global hyperparameter into a tweet-level decision.

Qualitatively, calibrated ACS behaves as expected. Self-contained tweets with confident local predictions receive high self-loop/root-local gates and low distant-edge gates. Sarcastic or stance-bearing replies receive higher gates to their parents and thread roots. This explains the improvement over the fixed-depth BioGraph variant and supports the paper's central claim: the model should select context, not merely aggregate it.

The calibration analysis is important because the ambiguity estimator is the weakest link in ACS: a raw local classifier can be over-confident on rare or sarcastic cases. Temperature scaling reduces ECE from 0.118 to 0.046 and improves macro-F1 from 0.824 to 0.829, while the random-gate sanity check falls to 0.807. Thus, the benefit of ACS comes from learned, calibrated routing rather than from stochastic neighbourhood dropout.

8.2 Why Does the GA Outperform PSO and Random Search?

The GA's advantage stems from the mixed discrete–continuous structure of the hyperparameter vector $\mathbf{X}$. PSO converges faster than GA, but velocity updates produce non-integer positions for L, H, d_h, and $d_{\max}$, which are then rounded and may break

exploitative momentum [31]. The GA avoids this by treating integer and continuous genes with separate mutation operators. The adaptive model also creates interactions between $d_{\max}$, τ_{ctx}, and r_{balance}: deeper graphs provide more candidate context, the gate threshold determines how much of it remains active, and balancing changes how minority-class nodes influence attention. This non-flat search surface explains why structured evolutionary search outperforms random sampling at equal budget.

8.3 When Does Conversational Context Help?

The random-graph ablation confirms that gains come from structured conversational topology rather than from the GNN inductive bias alone. Qualitative inspection reveals three patterns.

Sarcasm resolution. "Another day, another miracle cure." is classified as Very Positive by a flat BERTweet classifier but correctly labelled Very Negative once its parent tweet–a neutral news report about a retracted study–is selected by ACS and aggregated.

Stance propagation. A short reply that appears neutral in isolation inherits the negative polarity of the thread it rejects or endorses. ACS keeps the relevant parent/root edge while suppressing unrelated side replies.

Topic disambiguation. "It is unbelievable" is correctly classified as Positive in a vaccine-efficacy thread and Very Negative in a hospital-capacity thread because the selected root provides topical grounding.

8.4 Attention and Gate Specialisation

Figure 9 illustrates the mean attention weights assigned by each of the four heads in GATv2 Layer 1 across conversational edge types. Head 3 specialises on forward reply edges, suggesting that it encodes "what does this tweet reply to?". Head 2 distributes weight more evenly between root and self-loop edges, acting as a global-context reader. ACS complements this head specialisation: the gate decides whether an edge is reliable enough to enter the attention normalisation, while attention decides how strongly to weight the retained neighbours.

8.5 Limitations

Language coverage. The framework is evaluated on English tweets only. Multilingual extension requires language-specific or cross-lingual pretrained models and graph alignment.

API dependency and data volatility. Conversation graph reconstruction depends on the Twitter API. A non-trivial fraction of parent tweets is not recoverable, so future work should explore graph imputation for missing links.

Temporal bias. COVID-Emo is temporally distributed across pandemic phases, but the hydrated conversation graph depends on which tweets and parents remain accessible at hydration time. Temporal generalisation remains an open question.

Residual ACS uncertainty bias. Although the ambiguity head is calibrated with temperature scaling, calibration is not perfect. Some rare sarcastic cases may still be

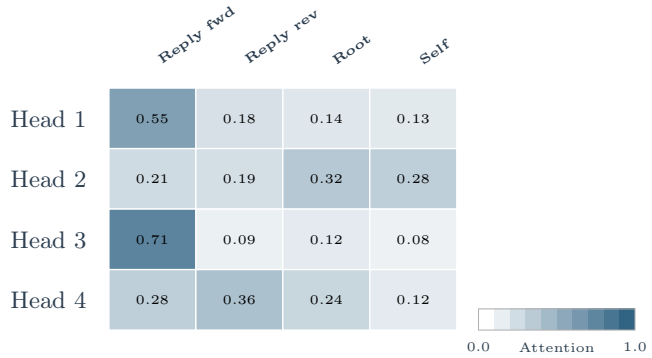


Fig. 9: Mean attention weight per head and edge type in GATv2 Layer 1 (darker = higher weight).

routed to insufficient context. Jointly calibrating the local uncertainty head and the graph encoder remains a future direction.

GA computational cost. Six hundred GA evaluations require roughly 5.8 h on an A100. The lightweight 300-evaluation configuration recovers most of the performance (0.823 vs. 0.829 macro-F1) at about half the optimisation time, making it the recommended setting for resource-constrained users. Warm-starting the GA from a short random-search phase may further reduce optimisation time, but this remains a future experimental direction.

Label quality. Some COVID-Emo tweets are ambiguous even for human annotators. Models trained on such data inherit this ceiling; macro-F1 near 0.83 may approach the practical annotation-consistency upper bound for this task.

8.6 Threats to Validity

Internal validity. Validation and test labels are never used for training. We further separate the calibration split $\mathcal{D}_{\text{cal}}$ from the validation split $\mathcal{D}_{\text{val}}$: the former is used only to fit the temperature parameter of the ACS ambiguity estimator, whereas the latter is reserved for GA fitness evaluation and early stopping. This avoids the double use of $\mathcal{D}_{\text{val}}$ for both calibration and model selection. Context-only parent/root nodes may appear in the graph, but they remain unlabelled and contribute only textual neighbourhood information.

External validity. Results are reported on COVID-Emo and SemEval-2017; other crisis domains may contain different conversational patterns, moderation policies, and class distributions. The SemEval-2017 experiment is explicitly separated into native training and supervised adaptation protocols, and should not be interpreted as zero-shot transfer.

Construct validity. The mapping from fine-grained emotions to ordered sentiment classes is necessarily approximate. We reduce this threat by reporting mapping-sensitivity and original-emotion evaluations, but mixed-affect tweets may still not be fully represented by a single ordinal sentiment label.

Statistical validity. We use 20 independent runs and paired tests, but ablation values should be interpreted as evidence for component usefulness rather than definitive population-level guarantees.

8.7 Broader Impact

Beyond COVID-19, BioGraph-SentiCOVID is applicable to domains where opinion mining benefits from conversational context, including customer service, political discourse, and mental-health related social media. These uses require safeguards: aggregate analysis is more appropriate than individual profiling, and deployment should include human review, bias monitoring, calibration checks, and transparent data governance. The adaptive context component can also reduce harmful over-interpretation by suppressing unreliable neighbourhood evidence rather than forcing every message into a broad thread context.

9 Conclusion

This paper presented **BioGraph-SentiCOVID**, a calibrated adaptive-context graph neural framework for COVID-19 tweet sentiment analysis. The framework models Twitter reply threads as directed multi-relational graphs, encodes nodes with BERTweet contextual embeddings, applies imbalance-aware learning, and uses a GA to tune the mixed hyperparameter space of the GATv2 classifier. Its main methodological novelty is **Calibrated Adaptive Context Selection**: a tweet-level and edge-level gating mechanism that uses temperature-scaled ambiguity, neighbour reliability, and hop distance to let each tweet exploit the amount of conversation context it needs rather than relying on one fixed graph depth for all instances.

Across 20 independent runs on COVID-Emo, BioGraph-SentiCOVID achieves a macro-F1 of 0.829, outperforming the fixed-depth BioGraph variant (0.817), GATv2-PSO (0.812 under the adaptive setup), and strong text-only baselines. Supervised SemEval-2017 adaptation yields macro-F1 = 0.851, while native SE17 training remains separately reported to avoid an unfair initialisation comparison. Ablation studies confirm that each component–contextual embeddings, real conversation topology, dynamic GATv2 attention, imbalance-aware learning, GA optimisation, calibrated ACS, and context-only nodes–contributes measurably to performance. Additional mapping-sensitivity and original-emotion evaluations show that the gains do not rely on one arbitrary emotion-to-sentiment mapping.

For practical use, the full GA configuration provides the best accuracy, whereas the lightweight 300-evaluation variant retains most of the performance at about half the optimisation cost. Future work will explore multilingual extension, heterogeneous graphs incorporating user/hashtag/URL nodes, joint calibration of the ambiguity head and graph encoder, warm-started or multi-objective GA optimisation, joint optimisation of BERTweet fine-tuning and graph hyperparameters, and temporal drift adaptation across pandemic phases.

The full codebase, re-hydration scripts, and optimal hyperparameter configurations will be released upon publication.

Declarations

Funding. Not applicable.

Conflict of interest. The author declares no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The experiments use publicly available benchmark datasets described in the manuscript.

Code availability. Code availability can be provided upon reasonable request.

References

[1] World Health Organization: Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. WHO Joint Statement (2020)

[2] Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 social media infodemic. Scientific Reports **10**(1), 16598 (2020) https://doi.org/10.1038/s41598-020-73510-5

[3] Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., Shah, Z.: Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. Journal of Medical Internet Research **22**(4), 19016 (2020) https://doi.org/10.2196/19016

[4] Sharma, K., Seo, S., Mack, C., Bhatt, S., Liu, Y.: COVID-19 on social media: Analyzing misinformation in Twitter conversations. arXiv preprint arXiv:2003.12309 (2020)

[5] Boon-Itt, S., Skunkan, Y.: Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. JMIR Public Health and Surveillance **6**(4), 21978 (2020) https://doi.org/10.2196/21978

[6] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: defend: Explainability-aware fake news detection. In: Proceedings of KDD, pp. 395–405 (2019). https://doi.org/10.1145/3292500.3330935

[7] Dou, Y., Shu, K., Xia, C., Yu, P.S., Sun, L.: User preference-aware fake news detection. In: Proceedings of SIGIR, pp. 2051–2055 (2021). https://doi.org/10.1145/3404835.3462990

[8] Sosea, T., Pham, C., Tekle, A., Caragea, C., Li, J.J.: Emotion analysis and detection during COVID-19. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6938–6947. European Language Resources Association, Marseille, France (2022). https://aclanthology.org/2022.lrec-1.750/

[9] Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018) https://doi.org/10.1016/j.neunet.2018.07.011

[10] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=rJXMpikCZ

[11] Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=F72ximsx7C1

[12] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. In: CS224N Project Report, Stanford University (2009)

[13] Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC, pp. 1320–1326 (2010)

[14] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 task 4: Sentiment analysis in Twitter. In: Proceedings of SemEval, pp. 1–18 (2016). https://doi.org/10.18653/v1/S16-1001

[15] Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proceedings of EMNLP, pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162

[16] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

[17] Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: A pre-trained language model for english tweets. In: Proceedings of EMNLP: System Demonstrations, pp. 9–14 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[19] Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. arXiv preprint arXiv:2005.07503 (2020)

[20] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive study on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems **32**(1), 4–24 (2021) https://doi.org/10.1109/TNNLS.2020.2978386

[21] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI Open **1**, 57–81 (2020) https://doi.org/10.1016/j.aiopen.2021.01.001

[22] Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of AAAI, vol. 33, pp. 7370–7377 (2019). https://doi.org/10.1609/aaai.v33i01.33017370

[23] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=SJU4ayYgl

[24] Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.: DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Proceedings of EMNLP, pp. 154–164 (2019). https://doi.org/10.18653/v1/D19-1015

[25] Ishiwatari, T., Yasuda, Y., Miyazaki, T., Goto, J.: Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proceedings of EMNLP, pp. 7360–7370 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.597

[26] Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, pp. 593–607 (2018). https://doi.org/10.1007/978-3-319-93417-4_38

[27] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**, 281–305 (2012)

[28] Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, vol. 25 (2012)

[29] Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, ??? (1989)

[30] Young, S.R., Rose, D.C., Karnowski, T.P., Lim, S.-H., Patton, R.M.: Optimizing deep learning hyper-parameters through an evolutionary algorithm. In: Workshop on Machine Learning in HPC Environments (MLHPC) (2015). https://doi.org/10.1145/2834892.2834896

[31] Lorenzo, P.R., Nalepa, J., Kawulok, M., Ramos, L.S., Pastor, J.R.: Particle swarm optimization for hyper-parameter selection in deep neural networks. In: Proceedings of GECCO, pp. 481–488 (2017). https://doi.org/10.1145/3071178.3071208

[32] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object

detection. In: Proceedings of ICCV, pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324

[33] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[34] Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of EMNLP, pp. 6382–6388 (2019). https://doi.org/10.18653/v1/D19-1670

[35] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 1321–1330 (2017). https://proceedings.mlr.press/v70/guo17a.html

[36] Li, G., Muller, M., Thabet, A., Ghanem, B.: DeepGCNs: Can GCNs go as deep as CNNs? In: Proceedings of ICCV, pp. 9267–9276 (2019). https://doi.org/10.1109/ICCV.2019.00936

[37] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7

[38] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=Skq89Scxx

[39] Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of SemEval, pp. 502–518 (2017). https://doi.org/10.18653/v1/S17-2088

[40] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of NAACL, pp. 1480–1489 (2016). https://doi.org/10.18653/v1/N16-1174

[41] Kennedy, J., Eberhart, R.: Particle swarm optimization. Proceedings of ICNN **4**, 1942–1948 (1995) https://doi.org/10.1109/ICNN.1995.488968

[42] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009) https://doi.org/10.1016/j.ipm.2009.03.002

[43] Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945) https://doi.org/10.2307/3001968

[44] Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds

(2019). arXiv:1903.02428

[45] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., *et al.*: Transformers: State-of-the-art natural language processing. In: Proceedings of EMNLP: System Demonstrations, pp. 38–45 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6