UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Detecting Fake News Spreaders on Twitter Using Transfer Learning

Diploma Thesis

Konstantinos Kokkalis

Supervisor: Dimitrios Rafailidis

September 2025

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Detecting Fake News Spreaders on Twitter Using Transfer Learning

Diploma Thesis

Konstantinos Kokkalis

Supervisor: Dimitrios Rafailidis

September 2025

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανίχνευση Χρηστών του Twitter που Διαδίδουν Ψευδείς Ειδήσεις με Μεταφορά Μάθησης

Διπλωματική Εργασία

Κωνσταντίνος Κόκκαλης

Επιβλέπων: Δημήτριος Ραφαηλίδης

Σεπτέμβριος 2025

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Approved by the Examination Committee:

Supervisor **Dimitrios Rafailidis**
Associate Proffesor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Dimitrios Katsaros**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Georgios Thanos**
Laboratory Teaching Staff, Department of Electrical and Computer Engineering, University of Thessaly

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dimitrios Rafailidis, for his continuous guidance and invaluable support throughout the course of my research. His expertise and encouragement have been instrumental in the completion of this work.

I would also like to extend my appreciation to Prof. Dimitrios Katsaros and Prof. Georgios Thanos, for kindly serving as reviewers of my thesis and for their valuable time and constructive comments.

Finally, I am deeply thankful to my family and close friends for their unwavering support, patience, and belief in me. Their encouragement and understanding have provided me with the strength to persevere during the most challenging stages of my studies.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Konstantinos Kokkalis

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Diploma Thesis

Detecting Fake News Spreaders on Twitter Using Transfer Learning

Konstantinos Kokkalis

Abstract

We study SeGA, a preference-aware self-contrastive graph model for anomalous user detection on Twitter. The model couples a heterogeneous graph encoder with prompt-driven supervision: large language models summarize each user's recent posts into topic-emotion preferences, which are converted into class prototypes and used to align user representations with a contrastive objective. Building on this idea, we contribute: (i) a principled downscaling of the TwBNT benchmark to a 10k-user subset that preserves class imbalance, edge heterogeneity, and degree distributions (ii) a lightweight model variant that removes one Relational Graph Transformer layer and shares parameters across relation types, cutting trainable parameters by 76.56% while stabilizing training on smaller graphs (iii) a comprehensive benchmark against five graph baselines and (iv) an exploratory transfer of SeGA to rumor veracity classification, including a stance-aware graph construction pipeline and a discussion of domain shift, stance reliability, and fine-tuning stability. Together, these results show that prompt-guided contrastive learning makes graph-based detection both data efficient and deployable, and that compact architectural choices can deliver strong precision in constrained environments without sacrificing interpretability.

Keywords:

Anomalous User Detection, Heterogeneous Graphs, Contrastive Learning, Prompts, Twitter, Misinformation, Model Compression

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Διπλωματική Εργασία

Ανίχνευση Χρηστών του Twitter που Διαδίδουν Ψευδείς Ειδήσεις με Μεταφορά Μάθησης

Κωνσταντίνος Κόκκαλης

Περίληψη

Μελετούμε ένα μοντέλο γράφων για την ανίχνευση ανώμαλων χρηστών στο Twitter. Το μοντέλο συνδυάζει έναν ετερογενή κωδικοποιητή γράφου με εποπτεία καθοδηγούμενη από προτροπές: μεγάλα γλωσσικά μοντέλα συνοψίζουν τις πρόσφατες αναρτήσεις κάθε χρήστη σε ζεύγη προτιμήσεων θέματος–συναισθήματος, τα οποία μετατρέπονται σε πρωτότυπα κλάσεων και χρησιμοποιούνται για την ευθυγράμμιση των αναπαραστάσεων των χρηστών μέσω αντιθετικού στόχου. Βασιζόμενοι σε αυτή την ιδέα, συνεισφέρουμε: (i) μια τεκμηριωμένη υποκλιμάκωση του TwBNT σε υποσύνολο 10 χιλιάδων χρηστών που διατηρεί την ανισορροπία κλάσεων, την ετερογένεια ακμών και τις κατανομές βαθμού, (ii) μια «ελαφριά» εκδοχή του μοντέλου που επιτυγχάνει μείωση των εκπαιδεύσιμων παραμέτρων κατά 76,56%, (iii) μια ολοκληρωμένη σύγκριση με πέντε μοντέλα ανίχνευσης ανώμαλων χρηστών και (iv) μια διερευνητική προσαρμογή του μοντέλου στη διάκριση αληθοφάνειας φημών, με κατασκευή γράφου με επίγνωση στάσης. Τέλος, συζητούμε τη μετατόπιση πεδίου, την αξιοπιστία της στάσης των χρηστών και τη σταθερότητα του fine-tuning. Συνολικά, τα αποτελέσματα δείχνουν ότι η αντιθετική μάθηση με καθοδήγηση προτροπών καθιστά την ανίχνευση σε γράφους ταυτόχρονα αποδοτική ως προς τα δεδομένα και αναπτύξιμη, ενώ οι συμπαγείς αρχιτεκτονικές επιλογές μπορούν να προσφέρουν υψηλή ακρίβεια σε περιορισμένα περιβάλλοντα χωρίς να θυσιάζεται η ερμηνευσιμότητα.

Λέξεις-κλειδιά:

Ανίχνευση Ανώμαλων Χρηστών, Ετερογενείς Γράφοι, Αντιθετική Μάθηση, Προτροπές, Παραπληροφόρηση, Twitter, Συμπίεση Μοντέλου

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Table of contents

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

List of figures

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

List of tables

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Abbreviations

RGT	Relational Graph Transformer
SSL	Self Supervised Learning
LLM	Large Language Model
HIN	Heterogenous Information Network
MLP	Multilayer Perceptron
InfoNCE	Information Noise Contrastive Estimation
GCN	Graph Convolutional Network
GAT	Graph Attention Network
GNN	Graph Neural Network
CLS	Classification Token
RGAT	Relational Graph Attention Network
BERT	Bidirectional Encoder Representations from Transformers
GPU	Graphics Processing Unit

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 1

Introduction

The rapid proliferation of social media has made platforms like Twitter central to real-time communication and public discourse. However, this ubiquity has also attracted malicious actors that manipulate public opinion, spread misinformation, and degrade online discussions ([2], [3]). Among anomalous account types, automated bots constitute the most prevalent category, representing between 9-15% of the active user base ([4]). The emergence of state-sponsored troll networks, exemplified by the 2,700 Russian-affiliated accounts disclosed by the U.S. Congress in 2017 ([5]), illustrates the evolving landscape of inauthentic user behavior.

Troll accounts are especially challenging to detect, as they are often operated by real individuals who mimic legitimate user behavior to evade conventional detection systems. As such, there is an urgent need for detection models that go beyond bot identification and can robustly classify a diverse set of anomalous users.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

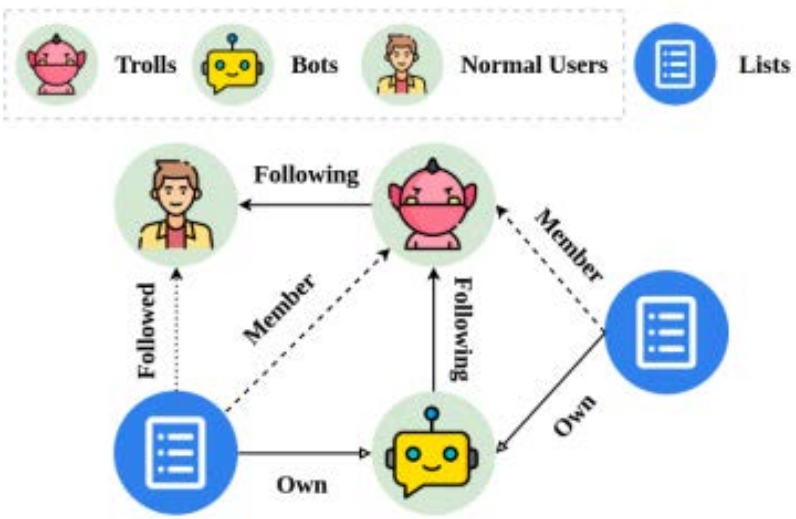

Figure 1.1: An example of anomalous user behavior in social media. Edges connecting entities represent diverse relationships[1].

Figure 1 illustrates how malicious actors on Twitter can exploit not only direct interactions with other users, but also the platform's list feature to amplify their deceptive campaigns. As an example scenario, a bot creates its own list and populates it with other anomalous accounts (such as trolls), thus establishing a network of coordinated misinformation. When an unsuspecting legitimate user subscribes to this list, they inadvertently receive the false or misleading content disseminated by its malicious members. This example underscores the challenge of distinguishing adversarial behaviors that mimic normal social activity, since troll accounts often evade detection by imitating human-like interaction patterns.

While prior research has demonstrated efficacy in bot detection through graph neural network architectures ([6],[7]) and heterogeneous information network analysis ([8], [9]), these approaches exhibit limitations when confronted with evolving deceptive strategies. The complexity of user interactions on Twitter, combined with the heterogeneous nature of its entities (e.g., users, tweets, lists), demands models that can exploit both relational and behavioral cues.

To address this, we explore and extend the SeGA framework: a graph-based model that integrates preference-aware self-contrastive learning with prompt-based representations. Our work applies, analyzes, and adapts this model to the new benchmark dataset (TwBNT), which includes automatically annotated bots, trolls, and normal users.

We begin with a detailed statistical analysis of the TwBNT dataset and develop a stratified downscaled version for efficient experimentation. The SeGA model is then adapted and eval-

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

uated under multiple configurations, including a modified lightweight variant. Experiments are conducted across both the original and downscaled datasets.

The remainder of this thesis is structured as follows: Chapter 2 establishes the theoretical foundation by examining existing literature in anomalous user detection and graph-based learning methodologies. Following this review, Chapter 3 introduces the TwBNT dataset, providing detailed insights into its organizational structure and the systematic downsampling approach employed. The core methodological contribution is presented in Chapter 4, which unveils the SeGA model and its preference-aware architecture. Chapter 5 details the implementation and architectural modifications to SeGA. Chapter 6 describes the experimental setup and evaluation metrics. Chapter 7 presents the results and discusses key findings and trade-offs. An exploratory investigation into SeGA's broader applicability is examined in the Appendix, which investigates the model's potential for transfer learning applications in rumor detection tasks. Chapter 8 concludes with a summary of contributions and directions for future research.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 2

Related Work

2.1 Bot and Troll Detection on Twitter

Social media platforms, and particularly Twitter, have become fertile ground for the spread of misinformation, propaganda, and coordinated influence campaigns. To address these threats, researchers have developed various automated detection methods targeting malicious users. Historically, the primary focus has been on detecting *bots*—automated accounts that generate content and interact with users at scale.

Early bot detection methods relied heavily on supervised learning using hand-crafted features extracted from user metadata, tweet content, and activity patterns [10, 11]. However, such approaches often failed to generalize across evolving bot behaviors.

In contrast, *troll detection* remains comparatively underexplored. Trolls are typically real users or sophisticated accounts designed to provoke, manipulate discourse, or influence public opinion. Their behaviors are less predictable and more nuanced, often involving emotionally charged or misleading content. Traditional classifiers struggle to detect trolls due to their similarity with normal users and the absence of explicit behavioral signals.

2.2 Graph-Based User Modeling

Graph-based learning has emerged as a powerful framework for modeling user interactions and relationships on social networks ([12], [13], [14]). Twitter, in particular, can be naturally represented as a heterogeneous information network (HIN), where nodes correspond to users or lists, and edges represent various relationships such as followership, mentions,

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

retweets, and list memberships [15], [16], [17].

Graph neural architectures considered here span early spectral methods to relation-aware transformers. **GCN** [6] formulates convolution in the spectral domain to propagate labels over the topology, but it largely assumes homogeneous edge types and node features. Addressing the uniform treatment of neighbors, **GAT** [7] introduces self-attention so that aggregation weights are learned from data. To cope with Twitter's multi-relational structure, **BotRGCN** [8] extends relational GCNs to distinguish interactions such as follows and mentions. Building on this line, **RGT** [18] combines explicit multi-relation encodings with transformer attention and reports strong results on TwiBot benchmarks. Finally, **SimpleHGN** [19] offers a lightweight edge-type attention mechanism that flexibly aggregates over heterogeneous relations without the full complexity of a transformer.

These models demonstrated significant improvements over traditional approaches by capturing both structural and semantic information embedded in user interactions. However, most of them treat anomaly detection as a binary classification task, focusing only on bots versus normal users, while neglecting finer distinctions such as troll behavior.

2.3 Self Supervised Learning (SSL)

Self-supervised learning (SSL) has emerged as a robust and effective methodology across various research domains, enabling the acquisition of contextualized representations through auxiliary pretext tasks, in fields such as natural language processing ([20]) and computer vision ([21],[22]). Within the SSL paradigm, predictive learning approaches have demonstrated considerable potential for automated account detection applications. For example, Feng et al.[9] leveraged follower metrics as self-supervised signals to augment bot detection performance. Nevertheless, utilizing singular features as self-supervised indicators fails to capture the heterogeneous nature of malicious user behavior and provides insufficient characterization of diverse user types.

2.4 Anomaly Detection and the TwiBot Benchmark

To support systematic evaluation, the TwiBot benchmark series introduced large-scale annotated datasets for bot detection. TwiBot-20 [23] and TwiBot-22 [24] represent users as

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

graphs and incorporate behavioral, linguistic, and relational data. Yet, they do not natively include troll annotations.

The **TwBNT dataset**, introduced in the context of the SeGA framework, extends TwiBot-22 with automatically labeled trolls. This advancement enables a more realistic anomaly detection scenario involving three classes: normal users, bots, and trolls. It provides a novel testbed for evaluating models under class imbalance and behavioral ambiguity.

2.5 The SeGA Model

The SeGA model [1] represents a significant innovation in the space of anomalous user detection. It introduces a preference-aware self-contrastive learning mechanism that generates prompt-based user representations. Unlike conventional supervised classifiers, SeGA constructs soft label embeddings for each user type using large language model (LLM) prompts and aligns user representations with their respective class prototypes.

SeGA's architecture combines:

- **A heterogeneous encoder** based on Relational Graph Transformers (RGT), capturing multitype edge relationships.
- **Prompt-based self-supervision**, creating semantically meaningful class prototypes via textual prompts.
- **Contrastive learning**, encouraging user embeddings to cluster near their respective prototypes while being distant from others.

SeGA has been shown to outperform all previously mentioned baselines in both binary and multi-class anomaly detection tasks. Its ability to integrate textual semantics, social structure, and prompt engineering represents a novel direction in graph-based detection models. Despite substantial progress in bot detection, existing methods still struggle to robustly identify trolls and to generalize across user behavior types. Prior models either require extensive labeled data or lack interpretability due to their black-box design. SeGA addresses many of these challenges, but remains computationally intensive and under-evaluated, on constrained or downsampled datasets.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 3

Dataset Analysis

3.1 Overview of the original TwBNT Dataset

The TwBNT dataset extends the TwiBot-22 benchmark by incorporating automated annotations for *troll users*, enabling a more comprehensive evaluation framework for anomalous user detection. It includes three user classes: normal, bot, and troll users, and captures various types of interactions in a heterogeneous graph structure.

Each node in TwBNT corresponds to a Twitter user or list, while edges represent social relationships, such as follower/following connections and list memberships. Following the same principles as earlier datasets, TwBNT models relational diversity through five distinct edge types and includes user-generated content in the form of tweets and descriptions. The statistics of the proposed TwBNT benchmark, can be found in Table 3.1 below.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Types	Item	Count	Total
2 node types ($\mathcal{A}$)	# user	100,001	120,789
	# list	20,788	
5 edge types ($\mathcal{R}$)	# following	751,927	1,312,929
	# followers	148,250	
	# membership	365,593	
	# followed	39,258	
	# own	7,901	
3 classes	# troll	896	100,001
	# bot	5,047	
	# normal	94,058	
3 splits	# train	70,000	100,001
	# valid	20,000	
	# test	10,001	

Table 3.1: Statistics of the original TwBNT dataset.

3.2 Class Distribution and Imbalance

A major characteristic of the TwBNT dataset is the extreme class imbalance:

- **Normal users:** 94,058 (94.06%)
- **Bot users:** 5,047 (5.05%)
- **Troll users:** 896 (0.89%)

This skew presents challenges for supervised classification models, which are often biased toward the majority class. As such, evaluation metrics like macro F1-score ([25], [26] are critical for fair model comparison.

3.3 Edge Types and Relational Structure

TwBNT encodes five distinct edge types:

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

- **Following (user → user):** 751,927 edges
- **Follower (user ← user):** 148,250 edges
- **Membership (user → list):** 365,593 edges
- **Own (user owns list):** 7,901 edges
- **Followed (list → list):** 39,258 edges

This relational diversity allows models to capture structural nuances in social interactions and community behavior. By quantifying each edge type, we identified how densely connected users were across different relational modalities. The high number of membership and following edges, in particular, indicated that users frequently interacted via lists and followerships, which are critical for community structure and information diffusion.

3.4 List Nodes and Ownership

An important aspect of the dataset is the presence of **list nodes**, which users can either *own* or be *members* of. Analysis of list ownership revealed that only a minority of users were responsible for list creation, with roughly 10% of the original user set acting as list owners. This skew presented challenges for downscaling, as a proportional reduction in list ownership would have either resulted in disproportionate structural losses, or complicated the graph topology of the downsampled subset.

3.5 Node Degree Distribution

User connectivity was analyzed across all edge types. Degree distributions exhibited a heavy-tailed pattern typical of real world social networks: a small subset of users (hubs) had very high degrees, while the majority had sparse connections. This insight guided the sampling strategy during dataset reduction to ensure representation across the degree spectrum and preserve graph topology.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

3.6 Dataset Downscaling Procedure

To enable experimentation on modest hardware, we implemented a deterministic downscaling procedure that preserves key statistical and structural properties of TwBNT.

3.6.1 Seed User Domain Selection

To preserve the original domain diversity in the downscaled 10k-user TwBNT dataset, we employed a seed-based stratified sampling approach rooted in high-level social domains. This method ensured that the sampled graph retained influential users from various thematic communities, thus maintaining semantic richness and topological variety.

We defined four high-level social domains: **Politics, Entertainment, Business, Sports**

For each domain, we curated a list of prominent Twitter accounts present in TwBNT, selected based on their public influence and platform visibility. These accounts are typically high-degree nodes in the social graph, meaning they have a large number of connections (followers, memberships) and often serve as central hubs in their respective communities. As such, they are well positioned to act as entry points for neighborhood expansion during downsampling. These accounts served as seed users. Example seed usernames include `@POTUS`, `@elonmusk`, `@taylorswift13`, and `@cristiano`.

Seed Selection Algorithm

To initiate the downscaling process, we implemented an exact-enforcement seed selection procedure. For each domain, the candidate seed usernames were matched in a case-insensitive way against the available user metadata. The goal was to select one seed user per domain: a scaled-down version of the 10 per domain strategy used in the full 100k-user dataset.

In cases where a direct match for a seed username was not found within a domain, we randomly sampled an alternative user from the same domain, prioritizing those labeled human. Care was taken to avoid duplicate selections across domains by maintaining a set of already-used user IDs. This approach resulted in a balanced set of four seed users, each representing a distinct thematic community.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Graph Integration and Validation

These seed users served as anchor nodes for the downscaled user graph. Their inclusion was verified by confirming their presence in the user node set and ensuring that they were well connected. For the remaining users, we used the same selection algorithm as the one used in TwiBot-20[23] (see Algorithm 1). Following their integration, we performed consistency checks to validate node and edge statistics, and to also confirm that no duplicates or sampling errors had occurred. This strategy enabled us to build a semantically diverse and structurally representative user graph, rooted in real-world community anchors drawn from politics, business, entertainment, and sports. The resulting graph contains a total of 10,000 users, categorized as follows:

- 9,406 normal users
- 504 bot users
- 90 troll users

Edge Filtering

We filtered each edge type to retain only those edges where both endpoints (in the case of user-user edges) or the user node (for user-list edges) belonged to the sampled user set. This ensured the consistency of network structure while eliminating dangling or orphaned nodes.

List Retention

TwBNT includes list nodes to which users may be connected via two relation types: ”membership” and ”own”. The former represents users being members of a list, while the latter represents users who have created or own a list. In the original dataset, approximately 10% of users were list owners. For the purposes of this downscaling process, we opted to retain only the membership edges. Users who were connected via ”own” edges (i.e., users who owned lists) were excluded from the downsampled subset. This decision was driven by the need to simplify the structural complexity of the graph while preserving the key statistical and topological properties of the original dataset. Incorporating list owners into the

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Algorithm 1: TwiBot-20/TwBNT-10k User Selection Strategy

Input : initial seed user u_0 in a user cluster

Output: user information set F

u_0.layer $\leftarrow 0$
$S \leftarrow \{u_0\}$
u_0.expanded $\leftarrow$ False
$F \leftarrow \varnothing$
while $S \neq \varnothing$ **do**
  $u \leftarrow S.\text{pop}()$
  $T(u) \leftarrow \text{get\_tweet}(u)$
  $P(u) \leftarrow \text{get\_property}(u)$
  if u.*layer* ≥ 3 ***or*** u.*expanded* $=$ True **then**
    $F \leftarrow F \cup \{\, u(T, P, N = \varnothing)\}$
  u.expanded $\leftarrow$ True
  $N^f(u) \leftarrow \text{get\_friend}(u)$
  $N^t(u) \leftarrow \text{get\_follower}(u)$
  $N(u) \leftarrow N^f(u) \cup N^t(u)$
  $F \leftarrow F \cup \{\, u(T, P, N)\}$
  $S \leftarrow S \cup N^f(u) \cup N^t(u)$
  foreach $y \in N^f(u) \cup N^t(u)$ **do**
    y.expanded $\leftarrow$ False
    y.layer $\leftarrow u$.layer $+ 1$
return F

10k-user subset posed a challenge, as preserving both the proportion of owners and their associated structural roles (e.g., list creation patterns) would have introduced significant imbalance and noise. Downscaling such users without distorting the relational structure or inflating the subgraph was not feasible. Therefore, excluding list ownership edges offered a principled trade-off that helped retain the heterogeneous node structure required by SeGA, without compromising dataset integrity or introducing bias.After filtering edges, we identified all list IDs still connected to the retained 10k users through membership edges, and extracted their corresponding metadata. This ensures that list nodes remain part of the heterogeneous information

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

network, allowing downstream models to leverage both user and list interactions effectively.

Tweet Filtering

We retained only tweets authored by the 10k users, ensuring that all auxiliary data aligns with the reduced user graph.

The table below summarizes the properties preserved during downscaling:

Property	**Preservation Method**
Class distribution	Stratified sampling by label frequency
Graph structure	Filtered edges within user subset, close overlap to the original structure (82.79% overlap ratio)
Node heterogeneity	Users and lists retained
Edge diversity	All relation types preserved except “own”
Schema compatibility	Output structure mirrors original
Reproducibility	Diverse domain seed sampling

Table 3.2: Preserved dataset properties in 10k-user TwBNT subset

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 4

The SeGA Model

4.1 Introduction

SeGA is a graph based model for detecting anomalous users on Twitter that jointly exploits heterogeneous relational structure and prompt-driven contrastive learning. Unlike purely supervised GNN classifiers (e.g., GCN, GAT), SeGA injects semantics of the labels and user behaviours (bot / troll / normal; topical and affective preferences) into the representation space via natural-language prompts. These prompts are encoded as vectors and used to anchor node embeddings with a contrastive objective, increasing separability even under label sparsity and class imbalance.

Two supervision signals are used:

1. **User-level pseudo-preference prompts** derived from a user's recent posts (topic, emotion).

2. **Class-level label prompts** describing the target classes (*bot*, *troll*, *normal*).

The first provides dense, self-supervised signals aligned with behavioural regularities; the second injects task semantics and improves calibration of the decision boundary. A lightweight classifier is trained on top of the contrastively aligned embeddings.

4.2 Architecture Overview

Components. SeGA comprises four modules:

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

1. **Heterogeneous Graph Encoder** (Relational Graph Transformer, RGT): relation-aware message passing over a Twitter HIN.

2. **Prompt Generation & Encoding**: LLM-generated user pseudo-preferences and hand-crafted (or LLM-curated) class descriptions, encoded as text embeddings.

3. **Self-Contrastive Objectives**: InfoNCE losses aligning node embeddings to (i) user-level pseudo-preference vectors and (ii) class-level prototypes.

4. **Classifier Head**: a linear/MLP head for final label prediction, trained jointly with cross-entropy.

Figure 4.1 summarizes the end-to-end flow.

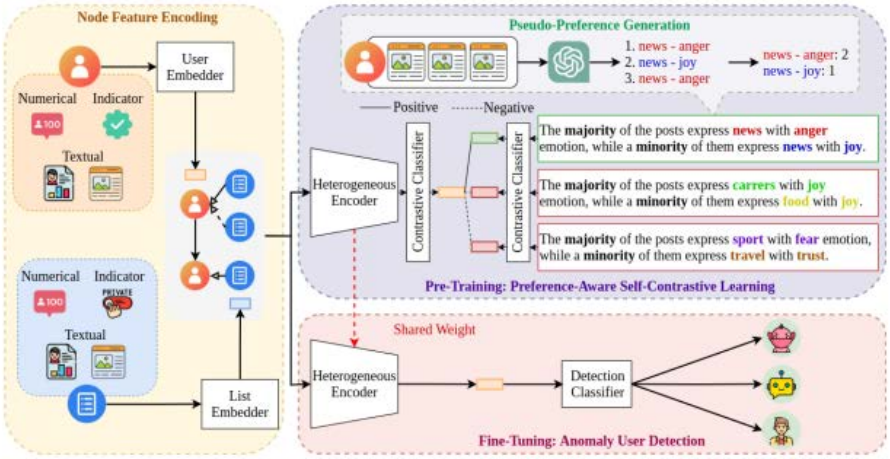

Figure 4.1: SeGA: heterogeneous graph encoding, prompt encoding, dual contrastive alignment (user-level and class-level), and supervised prediction [1].

4.3 Formal Problem Setup

We can model Twitter as a heterogeneous information network (HIN)

$$\mathcal{G} = \big(V, E, \mathcal{A}, \mathcal{R}, \phi, \psi\big),$$

where V are nodes (users u and lists l), E are typed edges, $\mathcal{A} = \{u, l\}$ are node types, $\mathcal{R}$ are relation types (e.g., `follow`, `follower`, `membership`, `followed`, `own`), $\phi : V \to \mathcal{A}$ maps nodes to types, and $\psi : E \to \mathcal{R}$ maps edges to relations. Let $\mathcal{N}_r(i)$ denote the neighbors of node i under relation r. The goal is to learn $f_\theta : \mathcal{G} \to \hat{y}_i \in \{\text{bot}, \text{troll}, \text{normal}\}$ for user nodes.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

4.4 Node Feature Construction

Each user (and list) node aggregates three feature families into a unified vector $\mathbf{x}_i$:

- **Indicator (binary) features** $\mathbf{x}_i^{\text{ind}} \in \{0,1\}^k$: e.g., verified flag, default profile image.
- **Numerical features** $\mathbf{x}_i^{\text{num}} \in \mathbb{R}^m$: e.g., tweet count, follower/following ratio; standardized (z-score).
- **Textual embeddings**: profile description and recent tweets encoded with SimCSE-RoBERTa [20].

We apply linear projections and non-linearities to each family and concatenate:

$$\tilde{\mathbf{x}}_i^{(\cdot)} = \sigma\big(W^{(\cdot)}\mathbf{x}_i^{(\cdot)} + \mathbf{b}^{(\cdot)}\big), \quad \mathbf{x}_i = \big[\tilde{\mathbf{x}}_i^{\text{ind}} \,\|\, \tilde{\mathbf{x}}_i^{\text{num}} \,\|\, \tilde{\mathbf{x}}_i^{\text{text}}\big],$$

with σ a LeakyReLU. For list nodes, we analogously embed list name/description and member statistics.

4.5 Relational Graph Transformer (RGT)

We employ a relation-aware multi-head attention layer to propagate information:

$$\mathbf{h}_i^{(\ell+1)} = \mathrm{LN}\Big(\mathbf{h}_i^{(\ell)} + \sum_{r\in\mathcal{R}} \sum_{h=1}^{H} \sum_{j\in\mathcal{N}_r(i)} \alpha_{ijrh}^{(\ell)}\, \mathbf{V}_{rh}^{(\ell)}\, \mathbf{h}_j^{(\ell)}\Big),$$

where $\mathbf{h}_i^{(0)} = \mathbf{x}_i$, H is the number of heads, LN is LayerNorm, and the attention weights are

$$\alpha_{ijrh}^{(\ell)} = \mathrm{softmax}_{j\in\mathcal{N}_r(i)}\left(\frac{\big(\mathbf{Q}_{rh}^{(\ell)}\mathbf{h}_i^{(\ell)}\big)^\top \big(\mathbf{K}_{rh}^{(\ell)}\mathbf{h}_j^{(\ell)}\big)}{\sqrt{d/H}}\right).$$

$\mathbf{Q}_{rh}^{(\ell)}, \mathbf{K}_{rh}^{(\ell)}, \mathbf{V}_{rh}^{(\ell)}$ are relation- and head-specific projections; d is the hidden size. A position-wise FFN with residual is applied after attention (Pre-LN transformer). Stacking L layers enables multi-hop reasoning over typed edges. In **original SeGA**, L=2; in our **modified** variant (Chapter 5), L=1 to reduce compute with small accuracy impact.

Complexity. Per layer: $O\big(H\cdot(|E|d + |V|d^2)\big)$. Memory scales as $O(H|E| + |V|d)$; we use dropout on attention and FFN, and gradient clipping for stability.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

4.6 Pseudo-Preference Generation (User-level Prompts)

LLM annotation. For each user i, we collect the T most recent tweets (we use $T{=}10$) and query an LLM [27] to assign a topic t_{ij} and emotion e_{ij} per tweet j. Topics are chosen from a 16-category inventory tuned to Twitter [28]; emotions follow Plutchik [29].

Preference extraction. We compute frequencies over topic–emotion pairs and identify the most and least frequent pairs

$$(\hat{t}_i^{\text{maj}}, \hat{e}_i^{\text{maj}}) = \arg\max_{(t,e)} f_i(t,e), \quad (\hat{t}_i^{\text{min}}, \hat{e}_i^{\text{min}}) = \arg\min_{(t,e)} f_i(t,e).$$

We render a short natural-language summary (prompt) per user:

> *"The majority of the posts express $\hat{t}_i^{maj}$ with $\hat{e}_i^{maj}$ emotion, while a minority of them express $\hat{t}_i^{min}$ with $\hat{e}_i^{min}$."*

This sentence is encoded via a sentence encoder (SimCSE-RoBERTa [20]) to produce the *user-level pseudo-preference vector* $\mathbf{p}_i^{\text{pref}} \in \mathbb{R}^{d_p}$. We apply a linear map to match the graph embedding space: $\tilde{\mathbf{p}}_i^{\text{pref}} = W_p \mathbf{p}_i^{\text{pref}}$. This aligns with recent practice of using LLMs as weak annotators via prompting/instructions [30], [31], [32].

4.7 Self-Contrastive Learning

Let $\mathbf{u}_i$ be the output embedding for user i from the RGT. We use cosine similarity $\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|}$ and temperature τ.

User-level (preference) contrast. For a mini-batch $\mathcal{B}$ of users, define positives $\tilde{\mathbf{p}}_i^{\text{pref}}$ and negatives $\{\tilde{\mathbf{p}}_j^{\text{pref}}\}_{j \neq i}$. The InfoNCE [33] loss is

$$\mathcal{L}_{\text{pref}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp\left(\text{sim}(\mathbf{u}_i, \tilde{\mathbf{p}}_i^{\text{pref}})/\tau\right)}{\sum_{j \in \mathcal{B}} \exp\left(\text{sim}(\mathbf{u}_i, \tilde{\mathbf{p}}_j^{\text{pref}})/\tau\right)}.$$

Class-level contrast. When class prototypes are used, we align $\mathbf{u}_i$ to its class c_i:

$$\mathcal{L}_{\text{cls-ctr}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp\left(\text{sim}(\mathbf{u}_i, \mathbf{p}_{c_i}^{\text{cls}})/\tau\right)}{\sum_{c} \exp\left(\text{sim}(\mathbf{u}_i, \mathbf{p}_{c}^{\text{cls}})/\tau\right)}.$$

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Supervised prediction. A linear/MLP head yields logits $\mathbf{z}_i = W_o\mathbf{u}_i + \mathbf{b}_o$ and probability $\hat{\mathbf{y}}_i = \mathrm{softmax}(\mathbf{z}_i)$. With ground-truth one-hot $\mathbf{y}_i$:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbf{y}_i^\top \log \hat{\mathbf{y}}_i.$$

Total objective. We jointly optimize

$$\mathcal{L} = \lambda_{\text{pref}}\mathcal{L}_{\text{pref}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls-ctr}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{reg}}\|\theta\|_2^2,$$

with λ's tuned on validation. If class labels are scarce, λ_{CE} can be reduced and the model still benefits from $\mathcal{L}_{\text{pref}}$.

SeGA augments a relation-aware GNN with semantically grounded, dual contrastive objectives. User-level prompts distill behavioural regularities into vectors, while class-level prompts inject task semantics. The RGT aligns these signals with the network structure, and a small classifier produces final predictions. This combination yields a representation space where anomalous users are more readily separable, even with limited labels and skewed class distributions.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 5

Proposed Model

5.1 Original Implementation

The original SeGA model was implemented according to the specifications in [1]: The input features for user and list nodes consist of two types: indicator and numerical features. The number of indicator features k is 3 for user nodes and 1 for list nodes. The number of numerical features m is 5 for user nodes and 4 for list nodes. All numerical features $\mathbf{N}$ are standardized using z-score normalization. The dimensions of the encoded text embeddings are as follows: the description embedding dimension d_{des}, tweet embedding dimension d_{twe}, and prompt embedding dimension d_p are each set to 768. The hidden layer dimension d_h is set to 32, and the output dimension d_{out} is 128. The dimensions of the user and list node feature aggregators d_u and d_a are both set to 64. The number of layers g in the Relational Graph Transformer (RGT) is set to 2. Following [18], the maximum number of tweets q retained for each user and list is 20, representing the most recent tweets. The maximum lengths for the user description s and each tweet L are set to 50 tokens, and zero-padding is applied for shorter sequences. During the fine-tuning stage, the loss weighting parameter λ is set to 3×10^{-5}. The model is trained for 100 epochs during pre-training and 150 epochs during fine-tuning. A dropout rate of 0.3 is applied throughout the network. The model is optimized using the AdamW optimizer [34] with a learning rate of 0.001 and a batch size of 2048.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

5.2 Modifications for a Lightweight Model

Given the significant computational resources required for training the original SeGA model on the entire 100k-user TwBNT dataset, we introduced targeted architectural modifications to improve computational efficiency, particularly for the smaller, downsampled 10k-user subset. These adjustments sought to reduce the computational complexity of the model, while preserving its effectiveness in anomalous user detection.

One major architectural change was the reduction of the number of layers in the Relational Graph Transformer (RGT). Prior work shows simplified GNNs can remain competitive and improve efficiency [35], [36]. In our case, we decreased the depth of the RGT from two layers in the original implementation to a single layer. While this modification inevitably limited the model's multi-hop reasoning capability, potentially constraining its ability to capture more complex, indirect relationships within the graph, it significantly improved computational efficiency and memory management, making the model far more practical for rapid iteration and experimentation on smaller graphs.

We also addressed parameter redundancy in the initial SeGA architecture. Originally, the model maintained separate Transformer-based convolutional layers (TransformerConv) for each distinct edge type in the heterogeneous graph. Recognizing that this design considerably increased model complexity, we consolidated these layers into a single shared TransformerConv across all edge types. This shared parameter approach greatly reduced the total number of trainable parameters without significantly compromising model accuracy or expressiveness, given the relatively homogeneous nature of user interactions on Twitter.

In addition to these structural simplifications, we fine-tuned key hyperparameters to further enhance computational performance. We reduced the number of transformation heads (`trans_head`) from two to one, simplifying the attention mechanisms within the transformer module. Moreover, we lowered the output dimension (`out_channel`) to 128, further reducing the memory footprint and potential for overfitting, especially given the limited size of the downsampled training set.

To fully leverage the available data without incurring prohibitive training costs, we adopted a two-stage pretraining and fine-tuning approach. Initially, the full SeGA model was pretrained on the complete 100k-user dataset, capturing large-scale structural and semantic patterns across the social graph. Subsequently, we finetuned the pretrained model on the downscaled 10k-user dataset. This approach allowed the lightweight model to retain the global

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

insights learned from the larger dataset, while efficiently adapting to the more resource-constrained scenario.

In our experimental evaluations, we explored and compared four distinct configurations of the SeGA model. These included the original full-scale model applied to both the full and downsampled datasets (referred to as "SeGA (100k, full)" and "SeGA (10k, full)" respectively), as well as the modified, lightweight version tested similarly ("SeGA (100k, modified)" and "SeGA (10k, modified)").

A quantitative comparison of trainable parameters and memory requirements highlights the impact of our modifications. The original SeGA model contained approximately 1.6 million parameters, corresponding to a total size of 3164 MB. After applying our architectural and hyperparameter adjustments, the modified SeGA model required only around 375 thousand parameters, reducing the model size to 752 MB. This represents a substantial parameter and memory reduction of approximately 76% compared to the original configuration (see Table 5.1).

Metric	**Full Model**	**Modified Model**	**Reduction (%)**
Trainable Parameters	1.6M	375K	76.56%
Model Size (MB)	3164	752	76.22%

Table 5.1: Trainable parameters and memory usage before and after architectural simplifications

These modifications collectively underscore a critical trade-off between model complexity and practical usability. By simplifying the architecture, we considerably improved computational efficiency and scalability. Although some multi-hop reasoning capabilities were inevitably diminished, our empirical results indicate that the modified model retains strong performance metrics, making it highly suitable for real-world applications where computational resources are limited or quick iterative experimentation is required.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 6

Experimental Setup

6.1 Objectives and Evaluation Strategy

The primary objective of our experiments is to evaluate the performance of the SeGA model under various architectural configurations and dataset sizes. We compare both the original and modified SeGA variants against a suite of strong graph-based baseline models on the TwBNT dataset.

To ensure fair comparisons across highly imbalanced classes (normal, bot, troll), we adopt evaluation metrics that reflect per-class performance and not just overall accuracy.

6.2 Evaluation Metrics

We report the following standard metrics:

- **Macro F1-Score:** The unweighted average F1 across all classes. Especially important for imbalanced classification.
- **Precision:** The proportion of correctly predicted positive samples out of all predicted positives.
- **Recall:** The proportion of correctly predicted positive samples out of all actual positives.

Each experiment was repeated with 3 random seeds, and we report the mean and standard deviation for each metric.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

6.3 Baseline Models

We compare SeGA to five widely used GNN architectures, all adapted for heterogeneous user graphs:

- **GCN** [6]: A foundational spectral graph convolution model.
- **GAT** [7]: Introduces attention mechanisms for neighbor aggregation.
- **SimpleHGN** [19]: Employs edge-type attention on heterogeneous graphs.
- **BotRGCN** [8]: Designed for bot detection with type-aware relational encoding.
- **RGT** [18]: Incorporates multi-head attention across edge types with relational encoding.

All models were trained both on the original and the downsampled 10k-user TwBNT subset, while using the same input features and splits to ensure a consistent comparison. Experiments were conducted on the original TwBNT dataset, as well as the downsampled subset, constructed as described in Chapter 3. The user set was split as follows: 70% training, 20% validation and 10% testing. Stratified sampling was used to preserve the class proportions across all splits.

6.4 Implementation Details

All experiments were performed on an Nvidia GTX 4060 Ti GPU, with 16 GB of VRAM, using PyTorch version 2.0.1. For experiments with the original SeGA model, the implementation was the same as described in Chapter 5. For the experiments with the modified SeGA model, implementation remained the same as the original, with the changes in hyperparameters as mentioned in Chapter 5. For the baselines, the implementation was kept the same as the tests performed in the original paper [1], based on the TwiBot-22 [24] guidelines. All models used the same node features (indicator + numerical + textual embeddings), and list nodes were retained via membership edges only. All dataset files, model checkpoints, and logs were version-controlled and stored using experiment tracking tools.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 7

Results and Discussion

7.1 Quantitative Results

The experimental results for all baseline models evaluated on the 10k-user TwBNT subset are summarized in Table 7.1. A comparison of the performance of the SeGA variants can be seen in Table 7.2. Each model was evaluated using macro F1-score, precision, and recall.

Model	**F1 Macro (%)**	**Precision (%)**	**Recall (%)**
GCN	45.37 ± 0.59	56.50 ± 3.40	42.26 ± 0.66
GAT	50.70 ± 5.98	60.57 ± 5.36	46.72 ± 5.27
BotRGCN	**59.02 ± 2.06**	**66.65 ± 1.58**	**56.93 ± 2.38**
RGT	57.58 ± 4.06	65.56 ± 1.70	54.99 ± 3.10
SimpleHGN	55.18 ± 3.90	63.82 ± 11.19	50.21 ± 2.68

Table 7.1: Performance metrics of baseline models on the 10k TwBNT dataset

Model Variant	**F1 Macro (%)**	**Precision (%)**	**Recall (%)**
SeGA (100k, full)	58.56 ± 1.24	65.27 ± 0.68	54.35 ± 1.85
SeGA (10k, full)	**60.14 ± 0.30**	65.12 ± 0.01	**57.03 ± 0.38**
SeGA (100k, modified)	52.12 ± 0.78	64.16 ± 0.03	51.20 ± 0.06
SeGA (10k, modified)	56.48 ± 0.03	**67.58 ± 0.02**	50.59 ± 0.03

Table 7.2: Performance of SeGA variants on both datasets

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

7.2 Analysis of Results

7.2.1 Performance Compared to Baseline Models

Across all experiments on the 10k-user TwBNT subset, SeGA—both in its original and modified forms—outperformed every baseline model evaluated. The strongest baseline was BotRGCN, achieving a macro F1-score of 59.02%, followed by RGT and SimpleHGN. These models leverage relational edge encoding or edge-type attention to capture the heterogeneous nature of Twitter graphs, and their performance confirms the effectiveness of relational reasoning in anomaly detection.

However, despite their relative strength, these baselines consistently underperformed compared to SeGA. The original SeGA model trained on the 10k subset achieved the highest macro F1-score (60.14%) and the best recall among all models. The modified SeGA variant, while slightly trailing in F1-score, surpassed all baselines in precision (67.58%), indicating that it makes fewer false-positive predictions than any other model tested.

GCN and GAT, which lack relation-type awareness, performed the worst, with macro F1-scores of 45.37% and 50.70%, respectively. These results reinforce the importance of modeling edge semantics in heterogeneous user graphs—something SeGA and the stronger baselines explicitly account for. Furthermore, while BotRGCN and RGT benefit from edge-type embeddings, their lack of semantic prompt alignment likely limits their ability to distinguish subtle behavioral differences between user classes, particularly in imbalanced scenarios.

Overall, SeGA's use of contrastive learning and prompt-based label prototypes provides it with a distinct advantage. This enables better generalization on minority classes (especially trolls), which the baselines struggle to classify correctly due to data imbalance and low semantic separability.

These findings suggest that the SeGA architecture, and particularly its contrastive learning and prompt-based components, offers a more robust and semantically informed representation of anomalous user behavior than traditional GNN-based models.

7.2.2 Precision–Recall Trade-Off

A notable trade-off emerged between the two SeGA configurations. The modified variant, designed for computational efficiency, exhibited a precision that was 2.46% higher than the full model. However, this gain came at the cost of a 6.44% reduction in recall. This trade-off

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

aligns with expectations: architectural simplifications reduce model capacity, which can lead to fewer true positives being detected — but the predictions that are made tend to be more reliable.

In applications such as misinformation mitigation or moderation automation, this kind of behavior may be preferable, as it prioritizes minimizing the false identification of benign users. Conversely, if the primary goal is to identify as many threats as possible (even at the risk of including some false positives), the full SeGA model is more appropriate.

7.2.3 Impact of Dataset Size and Structure

Interestingly, the full SeGA model yielded better performance on the 10k-user dataset than it did on the full 100k-user version. This counterintuitive result can be attributed to two key factors. First, the downsampled graph is structurally simpler: it excludes certain edge types (such as “own” edges linking users to lists) that are often sparse and introduce noise without significantly contributing to classification. Second, the sampling strategy employed during downscaling preserved class balance across the normal, bot, and troll categories. This likely enhanced the model’s ability to learn discriminative features for minority classes, particularly trolls, which are underrepresented in the original dataset.

7.2.4 Stability and Variance Across Runs

In terms of training consistency, the modified SeGA model showed minimal variance across different random seeds. For example, the standard deviation of macro F1-score was as low as 0.03. This suggests that the reduced architecture is less sensitive to initialization and training fluctuations, making it well-suited for constrained environments or repeated experiments where reproducibility is critical.

7.2.5 Case Study

Following the original paper[1], we sampled users who mainly expressed anger in their posts, in order to reflect the normal behavior of a user. We then visualize the t-SNE embedding results of our approach against two sample baselines: RGT, SimpleHGN. We also visualize the results for the modified SeGA model on the downsampled dataset, to help us understand the impact of our modifications on overall performance. The embeddings of RGT show am-

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

biguous collocation for groups of normal, bot, and troll users, while troll and bot users of SimpleHGN have more distinguishable distances but mix with normal users. Nonetheless, the proposed SeGA is able to describe different categories of users with clearer boundaries, which is attributed to preference-aware self-contrastive learning to describe user behaviors via the corresponding posts. Surprisingly, the modified version of SeGA is also able to distinguish different types of users with clear boundaries, although we can see that some normal users mix with bots (see 7.1).

a: SimpleHGN b: RGT c: Original SeGA d: Modified SeGA

Figure 7.1: Visualization of the user embeddings produced by the baselines and the SeGA variants. Normal users are pictured using green, bots with red and trolls with blue[1].

7.3 Discussion

In conclusion, the experimental results provide strong evidence for the efficacy of SeGA as an anomaly detection framework. The full SeGA model offers a balanced and comprehensive solution, excelling across all major metrics. The modified version, while simpler, remains highly competitive and is particularly valuable in precision-critical applications. Moreover, the results show that careful dataset curation — such as ensuring class balance and reducing edge-type noise — can significantly impact model performance, sometimes even more than raw dataset size. Lastly, the prompt-based and contrastive components of SeGA appear to play a key role in its ability to generalize under imbalanced and noisy real-world conditions.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Chapter 8

Conclusion and Future Work

This thesis examined SeGA as a practical and adaptable framework for anomalous user detection on Twitter, and showed how careful data engineering and targeted architectural simplifications can deliver strong performance without heavy compute. We began by dissecting and downscaling the TwBNT benchmark into a 10k-user subset that preserves the original graph's structural and statistical properties, enabling faster experiments while maintaining realism. Building on this foundation, we reproduced the original SeGA pipeline and introduced a lightweight variant that removes one RGT layer and shares parameters across edge types. The compact model reduces trainable parameters by 76.56% and model size by roughly three quarters, improving stability and wall-clock efficiency on small graphs with only a modest trade-off in recall. Across head-to-head comparisons with GCN, GAT, BotRGCN, RGT, and SimpleHGN, both SeGA variants achieved competitive or superior macro F1, precision, and recall; the lighter model tended to maximize precision (fewer false positives), while the full model retained stronger recall, a useful trade depending on downstream risk tolerance.

Beyond in-domain classification, we explored SeGA's transferability by adapting it to rumor veracity prediction on PHEME. This exercise clarified the real obstacles to cross-domain reuse: pretrained semantic and relational priors do not transfer unchanged when the label space, stance dynamics, and data scale shift abruptly. In particular, fixed stance thresholds and small labeled sets produced optimization instabilities and class collapse—issues that point to the need for domain-aware prompt design, conservative fine-tuning schedules, and stronger alignment objectives when moving from large, heterogeneous user graphs to smaller, conversation-centric datasets.

Several limitations temper these findings. A downsampled graph, even when statistically

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

faithful, cannot capture every higher-order dependency present at full scale. Our prompt formulations were intentionally simple and fixed; more expressive or adaptive prompts might yield further gains. And the transfer experiments, while informative, operated under tight data constraints, making conclusions about generalization necessarily tentative.

The most promising next steps follow directly from these observations. On the modeling side, a "medium" SeGA—reintroducing a lightweight second RGT layer or edge-type–adaptive capacity—could recover recall while keeping the footprint small. Prompt optimization via prompt tuning or instruction-tuned LMs, coupled with preference extraction that learns stance rather than thresholding cosine similarity, should strengthen the self-contrastive signal. On the data and training side, richer heterogeneity (retweets, mentions, quotes, temporal edges), label-efficient regimes (few-shot prototypes, adapters, or parameter-efficient fine-tuning), and simple graph/text augmentations can improve robustness. For deployment, streaming inference with incremental updates and drift detection would make SeGA viable in near–real time. Finally, systematic cross-platform studies (e.g., Reddit, YouTube) will test whether the combination of heterogeneous graph reasoning and prompt-driven alignment scales beyond Twitter.

In short, this work shows that SeGA's blend of heterogeneous relational modeling and prompt-informed contrastive learning can separate subtle, evolving behaviors with high accuracy, and that the same design can be made lightweight without losing its edge. With better prompt learning, modest capacity adjustments, and domain-aware transfer practices, SeGA (and similar models) can form a practical backbone for safeguarding public discourse across modern social platforms.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Bibliography

[1] Ying-Ying Chang, Wei-Yao Wang, and Wen-Chih Peng. Sega: Preference-aware self-contrastive learning with prompts for anomalous user detection on twitter. *AAAI Conference on Artificial Intelligence*, 2024. https://arxiv.org/abs/2110.11780.

[2] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 2016.

[3] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 2020.

[4] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 280–289. AAAI Press, 2017.

[5] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Proceedings of the Companion of the 2019 World Wide Web Conference (WWW)*, pages 218–226. ACM, 2019.

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2017.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2018.

[8] Shangsong Feng, Tong Zhou, Yuxiao Dong, Yizhou Yang, and Jie Tang. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 313–321, 2021.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

[9] Shang Feng, Han Wan, Ning Wang, Jian Li, and Ming Luo. SATAR: A self-supervised approach to twitter account representation learning and its application in bot detection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3808–3817. ACM, 2021.

[10] Clayton A Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[11] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[12] Fan-Yun Sun, Jordan T. Hoffman, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*, 2020.

[13] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

[14] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

[15] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, 2018.

[16] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *Proceedings of The Web Conference (WWW)*, 2019.

[17] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference (WWW)*, 2020.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

[18] Shangsong Feng, Tong Zhou, Yuxiao Dong, Yizhou Yang, and Jie Tang. Relational graph transformer for bot detection on twitter. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3789–3799, 2022.

[19] Jianxin Lv, Chuan Yu, Wenqi Ou, Zhanxing Zhang, and Qingqing Yang. Simplifying heterogeneous graph neural network with attention mechanism. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1579–1589, 2021.

[20] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910. Association for Computational Linguistics, 2021.

[21] Qilong Lv, Meng Ding, Qi Liu, Yaliang Chen, Wei Feng, Shourong He, Cheng Zhou, Jie Jiang, Yueming Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1150–1160. ACM, 2021.

[22] Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

[23] Lingfei Wu, Yizhou Yang, Shangsong Feng, and Jie Tang. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2725–2732, 2020.

[24] Shangsong Feng, Tong Zhou, Yuxiao Dong, Yizhou Yang, and Jie Tang. Twibot-22: Towards graph-based twitter bot detection. In *NeurIPS Datasets and Benchmarks Track*, 2022. `https://doi.org/10.1145/3534678.3539431`.

[25] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

[26] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 2015.

[27] Yiming Liu, Tianyi Han, Shizhe Ma, Jie Zhang, Yang Yang, Jing Tian, Hongshen He, An Li, Minghao He, Zheng Liu, Zhen Wu, Dongdong Zhu, Xinyi Li, Na Qiang, Dong Shen, Tong Liu, and Bao Ge. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *CoRR*, abs/2304.01852, 2023.

[28] Jim and Ann. Twitter topics – the mega list. `https://inboundfound.com/twitter-topics-list/`, 2020. Accessed: 2025-08-07.

[29] Bilal Ghanem, Davide Buscaldi, and Paolo Rosso. TexTrolls: Identifying russian trolls on twitter from a textual perspective. *CoRR*, abs/1910.01340, 2019.

[30] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[31] Jason Wei, Maarten Bosma, Vincent Zhao, et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.

[32] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in nlp. *ACM Computing Surveys*, 2023.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. OpenReview.net, `https://openreview.net/forum?id=Bkg6RiCqY7`.

[35] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[36] Xiangnan He, Kuan Deng, Xiaoyu Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[37] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[38] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[39] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.

[40] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[41] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017.

[42] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. PHEME dataset for Rumour Detection and Veracity Classification. `https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078`, June 2018. Accessed: 2025-08-08.

[43] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3402–3413, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

[44] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[45] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

Appendix

Transfer Learning with SeGA

1 Motivation

While SeGA was originally designed for bot and troll detection on Twitter, its architecture is well-suited for broader applications in social media analysis, especially tasks involving misinformation and rumor propagation. Our exploration focused on adapting SeGA for fake news detection through Twitter, evaluating three potential datasets: FakeNewsNet, PHEME, and LIAR.

2 Dataset Evaluation

FakeNewsNet [37],[38],[39] offers a comprehensive benchmark with news articles labeled as genuine or fake, accompanied by associated Twitter posts and user engagement metrics. The dataset's multimodal nature (text, social graphs, temporal data) makes it theoretically ideal for SeGA's graph-based approach. However, access limitations due to Twitter API requirements rendered it impractical for our study.

LIAR [40] provides a collection of 12.8K manually labeled political statements with fine-grained veracity ratings. While valuable for fact-checking research, its focus on short statements rather than conversational threads and lack of social network data made it incompatible with SeGA's architecture.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

PHEME [41] [42], [43] emerged as the optimal choice despite not being exclusively focused on fake news. Its structure of Twitter conversations during breaking news events, complete with veracity labels and reply networks, closely mirrors the TwBNT dataset used in SeGA's original training. The dataset covers five major events with annotations for both rumor/non-rumor classification and finer-grained veracity assessment (true, false, unverified).

3 Methodological Framework

Our methodology for preparing the PHEME dataset and implementing transfer learning with the SeGA model consists of four interconnected phases: data preprocessing, feature extraction, graph construction, and model fine-tuning. Each phase builds upon the previous one to create a robust pipeline for rumor veracity classification.

3.1 Data Preprocessing

The PHEME dataset contains source tweets (both rumors and non-rumors) along with their conversational replies and veracity annotations. We process the raw JSON files with careful error handling to account for encoding variations and malformed data. Each tweet receives two types of labels: a binary rumor classification label (0 for non-rumor, 1 for rumor) and a fine-grained veracity label (0 for true, 1 for false, 2 for unverified, and -1 for non-rumors which are excluded from veracity analysis). The reply structure is preserved by maintaining mappings between source tweets and their reactions, creating the foundation for subsequent graph construction.

3.2 Feature Extraction

We employ a dual-feature approach combining semantic and social signals. First, each tweet's textual content is encoded using RoBERTa-base, yielding 768-dimensional semantic embeddings extracted from the [CLS] token representation. These embeddings capture the core meaning of each tweet while being computationally efficient through batch processing with 32 samples per batch.

Complementing the semantic features, we derive stance features using a Sentence-BERT model (`paraphrase-MiniLM-L6-v2`) [44], [45], [46]. The model computes cosine sim-

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

ilarity between source tweets and their replies, classifying each reaction into one of three categories: supporting (similarity > 0.7), denying (similarity < 0.3), or neutral (intermediate values). For each source tweet, we aggregate these classifications into normalized 3-dimensional stance features representing the distribution of supporting, denying, and neutral reactions in its conversational thread.

The final feature representation concatenates both components, resulting in a 771-dimensional vector (768 from RoBERTa + 3 from stance features) for each source tweet. Reaction tweets receive zero-initialized feature vectors since their primary role is structural, rather than semantic.

3.3 Graph Construction

The conversational dynamics are formalized as a directed graph where nodes represent tweets and edges represent reply relationships. Each edge is typed according to the detected stance (0 for support, 1 for deny, or 2 for neutral) and made bidirectional to ensure information flow in both directions. This construction yields a rich representation where source tweets contain full feature information while reaction tweets primarily contribute to the graph topology. The resulting heterogeneous graph explicitly models both the conversational structure and the social dynamics of agreement and disagreement within rumor threads.

3.4 Dataset Partitioning

We create two distinct partitioning schemes to support different learning tasks. For general rumor detection, we maintain a balanced split with 1,285 samples each for validation and test sets, preserving the natural rumor-to-non-rumor ratio (approximately 1:1.67). The training set contains the remaining samples, constituting roughly 80% of the total data.

For veracity classification, we focus exclusively on rumor tweets (label = 1) and create a specialized validation set with 200 samples per veracity class (true, false, and unverified). This balanced validation set ensures fair evaluation across all veracity categories, while the test set maintains the natural class distribution to reflect real-world conditions.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

3.5 Transfer Learning with SeGA

The SeGA architecture processes our constructed graph through a Relational Graph Attention network (RGAT) layer with four attention heads. The model learns to propagate information across different edge types (support, deny, neutral) while applying LeakyReLU activation ($\alpha = 0.2$) for nonlinear transformations. A final linear layer maps the learned node representations to 3-class veracity predictions.

Training employs several techniques to handle class imbalance. We weight samples inversely proportional to the square root of their class frequencies:

$$w_c = \frac{\sqrt{1/(N_c + \epsilon)}}{\sum_{c'} \sqrt{1/(N_{c'} + \epsilon)}} \times 3 \tag{.1}$$

where N_c is the count of class c and $\epsilon = 10^{-8}$ provides numerical stability. Optimization uses the Adam optimizer with an initial learning rate of 10^{-4}, which automatically reduces by half when the validation F1 score plateaus for two consecutive epochs. Early stopping terminates training if no improvement occurs within ten epochs.

3.6 Evaluation Protocol

Our primary evaluation metric is the macro F1-score, which provides a balanced measure across classes. We supplement this with per-class precision and recall analysis to identify specific strengths and weaknesses. As a baseline comparison, we include results from a logistic regression classifier operating on the raw feature vectors, helping quantify the value added by the graph structure and attention mechanisms.

4 Integration Challenges of Pretrained SeGA with PHEME

The integration of pretrained SeGA model weights into the PHEME rumor classification pipeline revealed several significant challenges stemming from fundamental differences between the source and target domains. These issues primarily emerged from three key areas: domain mismatch between datasets, scale disparity in training data, and unstable optimization dynamics during fine-tuning.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

4.1 Domain Mismatch and Feature Noise

The pretrained SeGA embeddings, originally trained on the large-scale, multi-domain TwBNT dataset , proved suboptimal for PHEME's specialized rumor analysis task. This incompatibility manifested most noticeably in the semantic embedding space, where RoBERTa representations optimized for general social media discourse failed to capture the nuanced linguistic patterns characteristic of rumor conversations. The domain gap was particularly evident in handling speculative language patterns prevalent in unverified claims, where the pretrained embeddings tended to map similar rumor and non-rumor phrases closer together than appropriate for the veracity classification task.

A related issue emerged in the stance detection subsystem, where the fixed cosine similarity thresholds established during SeGA's pretraining (support > 0.7, deny < 0.3) failed to generalize to PHEME's distinctive reply semantics. The conversational dynamics in rumor threads often exhibited different interaction patterns compared to general social media exchanges. For instance, neutral replies in rumor discussions frequently contained higher lexical overlap than expected (e.g., ”Is this confirmed?” or ”Has this been verified?”), causing the stance classifier to mislabel them as supporting reactions. This misclassification propagated through the graph structure, corrupting the edge-type representations that are crucial for SeGA's relational reasoning.

4.2 Data Scale Disparity and Overfitting

The dramatic difference in dataset scales between SeGA's pretraining corpus and PHEME's annotated examples created severe optimization challenges. Where SeGA originally trained on orders of magnitude more examples, PHEME's veracity annotations provided only approximately 200 labeled instances per class. This extreme data scarcity led to several pathological behaviors during fine-tuning, most notably complete class collapse where the model learned to predict only the majority ”true rumor” class while ignoring the ”false” and ”unverified” categories.

The scale mismatch also affected the model's ability to maintain useful pretrained representations. With such limited training examples relative to the model's capacity, the fine-tuning process tended to overwrite generally useful features with PHEME-specific artifacts. This manifested in unstable validation metrics, where macro-F1 scores varied by ±0.15 across different random seeds, indicating that the model was learning dataset-specific noise rather

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

than generalizable patterns.

4.3 Optimization Instability

The transfer learning process revealed fundamental tensions in the optimization dynamics when applying large pretrained models to small specialized datasets. The learning rate settings that worked well for SeGA's original training proved too aggressive for PHEME fine-tuning, causing violent parameter updates that disrupted carefully learned feature extractors. This was exacerbated by imbalance in the gradient scales between the pretrained components and the newly initialized classification head, where updates to the final layers tended to dominate and erase useful pretrained knowledge.

These optimization challenges became visually apparent in the training curves, which showed wild oscillations in the loss values without corresponding improvements in validation metrics. Analysis of parameter histograms confirmed that the fine-tuning process was causing severe distributional shifts in the pretrained layers, effectively destroying the hierarchical feature representations that make transfer learning valuable.

4.4 Proposed Mitigation Strategies

To address these challenges, we recommend a multi-pronged adaptation strategy, for future use:

- **Embedding Space Alignment**: Employ contrastive learning objectives to better align RoBERTa's semantic space with PHEME's rumor-specific language patterns while preserving generally useful linguistic knowledge.
- **Stance System Refinement**: Replace the fixed threshold stance detector with a learned classifier trained on PHEME reply pairs, better capturing domain-specific interaction patterns.
- **Conservative Fine-Tuning**: Implement layer-wise learning rate reduction and gradient clipping to stabilize optimization while protecting pretrained features.
- **Data Augmentation**: Develop rumor-specific augmentation techniques, such as controlled edge dropout and synthetic stance variation, to artificially expand the effective training set.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14

- **Modular Adaptation**: Explore adapter-based fine-tuning approaches that introduce small, task-specific parameters while keeping the majority of pretrained weights frozen.
- **Dataset Change**: If possible. switch to FakeNewsNet, which is structured similarly to SeGA and provides a better starting point for experimentation.

These challenges highlight the often-underestimated complexity of transferring social media analysis models between different contexts and scales. Successful adaptation requires careful consideration of both the technical architecture and the fundamental differences in how language and social interactions manifest across domains. The solutions outlined here provide a pathway to preserve SeGA's powerful relational reasoning capabilities while making it effective for the specialized task of rumor veracity classification.

Institutional Repository - Library & Information Centre - University of Thessaly
12/09/2026 23:19:32 EEST - 216.73.216.14