Identifying Misinformation Spreaders through Multi-Source User Modeling

Roman Chervinskyy
Johannes Kepler University Linz
Linz, Austria
r.chervinskyy@gmail.com

Antonela Tommasel
Institute of Computational Perception
Johannes Kepler University Linz
Linz, Austria
ISISTAN
CONICET-UNCPBA
Tandil, Argentina
antonela.tommasel@jku.at

Abstract

Social media platforms have become a major channel for information consumption, but also facilitate the rapid spread of misinformation. While much prior work focuses on detecting misleading content, identifying *users* who are likely to spread misinformation remains a more challenging task, as such behavior depends not only on textual signals but also on interaction patterns and temporal dynamics. In this paper, we model misinformation spreader detection as a *user-level behavior prediction* problem and propose a hybrid architecture that combines content-derived user features with graph-based representations of user similarity, user interactions, and content interactions. Experimental results on a benchmark dataset show that integrating multiple user representations improves misinformation spreader detection. These findings highlight the potential of multi-source user representations for user-centered misinformation detection.

CCS Concepts

• **Computing methodologies** → **Machine learning algorithms**;
• **Information systems** → **Social networks**.

Keywords

Misinformation, Fake News, Linguistic analysis, User modeling

ACM Reference Format:
Roman Chervinskyy and Antonela Tommasel. 2026. Identifying Misinformation Spreaders through Multi-Source User Modeling. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3774935.3812721

1 Introduction

Online social media platforms have become central spaces for information consumption, with 86% of users in 2024 reporting that they accessed news via a digital devices at least occasionally[1]. At the same time, these platforms facilitate the rapid spread of false and misleading information, which has been shown to spread faster than accurate news online[2]. This raises challenges not only for content moderation but also for understanding the behavior of users who repeatedly contribute to such dissemination. While prior work has focused on detecting misinformation at the level of individual posts or claims, identifying *users* who are likely to spread misinformation remains a distinct and relevant problem. Then, user-level detection can support moderation, monitoring, and intervention strategies beyond the classification of isolated content.

Identifying misinformation spreaders is challenging. First, misleading content is often difficult to distinguish from legitimate information based on text alone, as it may include partially correct claims, references to real events, or information whose credibility is not immediately verifiable. In some cases, misinformation is also disseminated strategically as part of coordinated influence operations, such as those reported in the context of the 2016 United States election [17]. Second, misinformation spreading is inherently social, users interact through replies, mentions, and shared content, forming patterns that cannot be captured from isolated publications alone. Finally, user behavior evolves over time, meaning that effective detection requires modeling not only what users share, but also how their activity develops across publications and interactions.

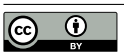
These challenges motivate a *user modeling* perspective on misinformation spreader detection. Rather than treating users as isolated collections of posts (e.g., [4, 6, 14, 15]), we represent them through multiple complementary views, including profile- and content-derived features, user interaction patterns, publication interactions, and similarity to others. Building on this, we propose a hybrid model that combines graph- and feature-based representations to predict whether a user is likely to spread misinformation. We evaluate the approach in a temporal setting, reflecting the challenge of generalizing to new accounts. Results show that integrating these multiple user representations improves performance over state-of-the-art approaches, highlighting the value of multi-source user modeling for identifying misinformation spreading behavior.

2 Related Work

Early approaches primarily relied on linguistic characteristics extracted from user-generated content [10]. These methods model users through textual and stylistic properties, assuming that deceptive behavior can be inferred from writing patterns alone. Examples

[1] https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/

[2] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308/



This work is licensed under a Creative Commons Attribution 4.0 International License.
UMAP '26, Gothenburg, Sweden
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3812721

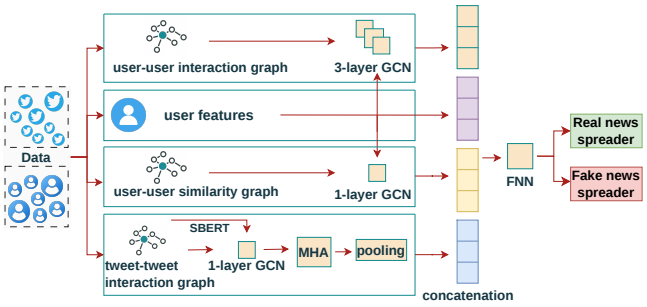


Figure 1: Pipeline overview of a proposed model

include models based on writing style, sentiment, readability, personality traits, and manually engineered user features [4]. While such methods capture important content signals, they overlook the relational context in which misinformation is shared.

To address this limitation, later work models spreading using graph-based representations, capturing not only what users post but also how they are connected and how information propagates. Rath et al. [16], for instance, incorporate interpersonal trust into graphs to guide neighborhood sampling and aggregation, while Hu et al. [7] combine textual representations with speaker-level credibility indicators. Although these methods better capture relational dependencies, they often rely on domain-specific assumptions or auxiliary signals that may not generalize well.

A third line of work combines content- and graph-based information into hybrid user representations. Hamdi et al. [6] combine Node2vec embeddings with user-level features, Qureshi et al. [15] derive graph-based features for downstream classification, and Tommasel et al. [18] integrate content and interaction graphs with textual and behavioral features. While these approaches highlight the value of combining multiple signals, they often rely on complex pipelines or heavyweight architectures, and typically underrepresent temporal dynamics. These limitations motivate our work, which models misinformation spreading as a multi-source user representation problem and integrates content, interaction, similarity, and temporal signals within a unified framework.

3 Multi-Source User Representation

To model misinformation spreading as a user-level behavior prediction task, we represent each user through multiple complementary views that capture different aspects of their online activity. Specifically, our approach combines: (1) a *User-with-User Interaction Graph* (UUIG) to capture social interactions, (2) a *User-with-User Similarity Graph* (UUSG) to model semantic similarity between users, (3) a *Publication-with-Publication Interaction Graph* (PPIG) to represent local conversational context, and (4) a set of user-derived features describing profile, linguistic, and behavioral characteristics. Figure 1 provides an overview of the proposed framework[3].

UUIG captures explicit interactions between users, such as replies, mentions, or responses, independent of the specific interaction type, which has proven effective in prior work [7, 16]. An edge is created whenever a user's publication directly involves another user. To complement these explicit social ties, we also construct a *UUSG*, following prior work on similarity-based user modeling [14]. In this graph, users are connected based on cosine similarity of *Sentence-Transformer* embeddings, with edges only when similarity exceeds a threshold Φ. To bound the influence of highly central nodes, each user is only connected to its top-k most similar neighbors.

To further incorporate contextual information, we model publication–conversation threads through the *PPIG*. Publications are encoded using BERT [2], and their contextualized representations are processed with a Graph Convolutional Network (GCN) [9] to capture dependencies between related publications. As user behavior unfolds over time, we additionally aggregate publications within short temporal windows. Publications shared within a one-day window are assumed to be semantically related, likely referring to the same event or topic, and are grouped accordingly to model the discussion cycle of a claim. Each group is weighted based on its size, while a time-based weighting mechanism assigns greater importance to more recent activity using a scaled function of the publication timestamps. The temporal weights are bounded to ensure that older publications still contribute to the final representation, preserving behavioral continuity.

Graph-based views are encoded using standard Graph Convolutional Networks (GCNs) [9] applied in parallel to the *UUIG*, *UUSG*, and *PPIG*. Graph components use varying depths, allowing the model to capture both local and broader relational patterns, including direct and indirect interactions between users. The resulting graph-based representations are concatenated with the full set of user-derived features to form a unified user representation. While all features are included in the final fusion step, a subset is also used as input to the graph encoders. This combined representation is then passed to a feed-forward neural network for final prediction.

4 Experimental Evaluation

Dataset. We evaluate our approach on the *FibVid* dataset [8], which contains social media posts associated with verified news claims collected during the COVID-19 period. Claims are labeled as authentic or fake based on fact-checking sources. The dataset comprises approximately 24,000 users and 296,000 posts, along with user metadata and interaction information.

Experimental settings. A user is defined as a misinformation spreader based on the proportion of their shared content labeled as misinformation. Let p denote the fraction of a user's publications labeled as misinformation. We consider two task formulations. In the binary setting, users with $p > 0.5$ are labeled as spreaders and the rest as non-spreaders. Under this threshold, 42% of users are labeled as non-spreaders and 57% as spreaders. To capture less clear-cut behavioral patterns, we also consider a three-class setting based on two thresholds, $k_1 = 0.35$ and $k_2 = 0.65$. Users with $p \leq k_1$ are labeled as non-spreaders, those with $p > k_2$ as spreaders, and those with $k_1 < p \leq k_2$ as potential spreaders. This intermediate class is intended to capture users whose behavior is ambiguous or unstable, and who may require further observation before a confident decision can be made. To evaluate generalization to previously unseen users, data are split at the user level using a 76/24 temporal train–test partition based on the date of each user's first observed interaction. The validation set is created by randomly sampling 14% of users from the training split. Each user is assigned to a single split, ensuring that training, validation, and test sets contain disjoint users. This setup reflects a realistic moderation scenario in which

[3]Code and data is at: https://github.com/hcai-mms/misinformation-spreaders-detection

systems must assess newly observed users based on their behavior. All models are trained and evaluated on identical data partitions to ensure comparability. We report Precision, Recall, F1-score, and Matthews Correlation Coefficient (MCC) [1], which is particularly suitable for imbalanced classification. In the three-class setting, Precision, Recall, and F1-score are computed using macro averaging. Given the practical importance of identifying misinformation spreaders, recall is treated as a particularly important metric.

Baselines. We compare the proposed model against a set of feature-based baselines that capture different aspects of user behavior. These baselines allow us to assess the contribution of individual behavioral signals as well as their combination. These include: (i) user profile statistics (user stats), (ii) linguistic features such as LIWC categories [13], readability metrics, punctuation usage, and tweet-content embeddings (*all-MiniLM-L6-v2* [19]), (iii) behavioral and interaction-based features such as tree metrics and content evolution, and (iv) higher-level user attributes including personality traits [11], emotion and sentiment features based on the NRC lexicon [12], and morality scores derived from Moral Foundations Theory [5]. We also consider a combined feature baseline that concatenates all feature groups. For each feature set, we evaluate a range of standard classifiers, including Gradient Boosting, Logistic Regression, Naive Bayes, Decision Tree, Feedforward Neural Network (FNN), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest (RF). The best-performing classifier is selected based on the Matthews Correlation Coefficient (MCC) on the validation set, with hyperparameters optimized via grid search. We further compare against state-of-the-art approaches, including Hamdi et al. [6], *CheckerOrSpreader* [4], Tommasel et al. [18], and *FacTweet* [3]. When available, original implementations are used, and hyperparameters are set following the respective studies.

Implementation Details. The proposed model and baselines were implemented in Python using *scikit-learn*, *PyTorch*, *TensorFlow*, and publicly available Hugging Face models[4]. User-derived features are normalized by clipping extreme values to the range $[-2, 2]$ to reduce noise [20]. The *UUIG* is encoded using three stacked GCN layers (size 32), the *UUSG* with a single GCN layer (size 32), and the *PPIG* with two GCN layers (sizes 64 and 16), followed by multi-head attention (two heads of size 32). The varying graph depths allow the model to capture information from neighborhoods of different ranges, with the *UUIG* incorporating both direct and indirect interactions between users. For each graph, the GCN produces contextualized node representations, and only the vector corresponding to the target user is retained for downstream classification. GCNs are applied in parallel to the *UUIG*, *UUSG*, and *PPIG*. The resulting graph-based representations are concatenated with the full set of user-derived features to form the final user representation. While all features are included in this final fusion step, a subset is also used as the input feature matrix X for the *UUIG* and *UUSG*. Tree metric features are used as graph input for these components. The final feed-forward classifier uses a hidden architecture of $[32, 64, 32]$ with ReLU activations. Training is performed using AdamW with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Hinge loss is used for binary classification, and training is conducted in mini-batches of 12 users for up to 13 epochs. For the *UUSG*, edges are created when cosine similarity exceeds a threshold of $\Phi = 0.67$, retaining the top-k most similar users per node ($k = 150$). Hyperparameters are selected based on validation performance via grid search. Experiments are conducted on an Acer Predator PH-315-52 equipped with an Intel Core i7-9750H CPU and an NVIDIA GeForce GTX 1660 Ti GPU. Training requires approximately 79 seconds per epoch, and inference takes approximately 0.06 seconds per user.

[4]https://huggingface.co/models

Classifier	F1-score	Precision	Recall	Matthews
Our proposal	**0.823**	0.794	<u>0.855</u>	**0.613**
all features	0.767	<u>0.801</u>	0.736	0.532
content evolution	<u>0.793</u>	0.687	**0.938**	0.515
LIWC categories	0.496	0.593	0.427	0.104
punctuation	0.535	0.721	0.425	0.262
readability	0.453	0.669	0.342	0.173
tree metrics	0.741	0.67	0.828	0.392
user-stats	0.412	0.693	0.293	0.178
emotion	0.251	0.559	0.162	0.028
personality	0.601	0.759	0.499	0.338
morality	0.229	0.596	0.141	0.052
tweet-content embedding	0.767	0.767	0.736	<u>0.538</u>
Tommasel et al. [18]	0.76	**0.806**	0.72	0.527
CheckerOrSpreader [4]	0.655	0.79	0.576	0.381
Hamdi et al. [6]	0.725	0.674	0.784	0.371
FacTweet [3]	0.416	0.656	0.304	0.148

Table 1: Performance for binary classification. Highest performance is in bold, and the second-best is <u>underlined</u>.

5 Experimental Results

Binary Classification. Table 1 reports the performance of all evaluated methods. The proposed model achieves the highest F1-score (0.823) and MCC (0.613) among all approaches, while also obtaining strong recall (0.855) and competitive precision (0.794). Overall, this indicates a well-balanced trade-off between precision and recall compared to the baselines. Bootstrapped comparisons further support this trend, with mean MCC improvements over the baselines ranging from 0.08 to 0.46 and 95% confidence intervals.

As it can be observed, the feature-only baseline using all user-derived features already performs competitively, outperforming most individual feature sets and several prior approaches. This suggests that different behavioral signals provide complementary information when combined. However, incorporating graph-based representations further improves performance, highlighting the importance of modeling relational and contextual aspects of user behavior. Among state-of-the-art methods, Hamdi et al. [6] achieve high recall with lower precision, whereas Tommasel et al. [18] attain high precision but lower recall, reflecting different trade-offs between these metrics. In contrast, the proposed model achieves a more balanced performance across evaluation measures.

Multi-Class Classification. Table 2 reports the results for the 3-class setting. To enable comparison, state-of-the-art baselines were adapted to this scenario using cross-entropy loss. The proposed model achieves the highest F1-score, recall, and MCC, while ranking second in precision behind the feature-based baseline using all attributes. To better understand the role of the intermediate class, we examine the corresponding binary setting obtained by excluding users whose misinformation ratio falls within the gray interval $[0.35, 0.65)$. In this setting, the model achieves a precision of 0.775 and a recall of 0.942. The gray interval contains 559

Classifier	F1-score	Precision	Recall	Matthews
Our proposal	**0.659**	0.692	**0.645**	**0.542**
all features	0.622	**0.719**	0.602	0.525
content evolution	0.561	0.69	0.543	0.408
LIWC categories	0.295	0.352	0.365	0.105
punctuation	0.404	0.39	0.431	0.247
readability	0.373	0.377	0.409	0.198
tree metrics	0.463	0.557	0.463	0.301
user-stats	0.339	0.361	0.384	0.143
emotion	0.325	0.374	0.364	0.081
personality	0.437	0.418	0.465	0.329
morality	0.279	0.39	0.355	0.052
tweet-content embedding	0.492	0.467	0.52	0.469
Tommasel et al. [18]	0.588	0.616	0.584	0.41
CheckerOrSpreader [4]	0.521	0.526	0.52	0.404
Hamdi et al. [6]	0.403	0.391	0.428	0.265
FacTweet [3]	0.36	0.346	0.383	0.127

Table 2: Performance for 3-class classification. Highest performance is in bold, and the second-best is underlined.

users, of which 86% are non-spreaders and 14% are spreaders, suggesting that this region is dominated by users whose behavior is closer to non-spreading but remains insufficiently clear-cut for confident classification. As expected, excluding this intermediate class slightly improves performance by reducing task ambiguity. Still, the multi-class formulation remains informative by explicitly capturing borderline cases that may benefit from delayed or cautious classification, particularly in moderation or monitoring scenarios where additional observations over time can help clarify behavior.

6 Conclusions

We proposed a multi-source user modeling approach for misinformation spreader detection combining graph-based representations with user behavioral features. Results show improved performance over prior approaches in both binary and multi-class settings. Preliminary ablation experiments suggest that model components are complementary, supporting the value of combining interaction-, similarity-, and content-based user representations[5]. Findings highlight the importance of modeling misinformation spreading through a richer view of user behavior and context.

Several directions for future work emerge. Evaluating the approach on topics beyond COVID-19 would help assess its robustness across domains and narratives. Exploring online and adaptive learning settings, where user behavior evolves over time and new data becomes continuously available, could improve the model's ability to capture shifting patterns of spreading. Another promising direction is the development of lighter (and more efficient) variants of the architecture. In preliminary experiments, we explored a simplified version of the model that merged the user interaction and similarity graphs into a single adjacency structure and applies a single GCN. Early results suggested that this simplification can reduce computational cost by a 40%, while maintaining competitive predictive performance. Modeling users' exposure to misinformation is another important direction, as observed behavior may reflect both exposure and individual propensity. Lastly, the increasing availability of LLMs opens opportunities to generate higher-level user representations, summarize behavioral patterns, or create auxiliary features that complement graph-based and linguistic signals.

[5]Preliminary results for the ablation study can be found in the repository

Ethical Considerations. Predictions should not be used as fully automated tools for punitive moderation, but rather as decision-support mechanisms for monitoring, prioritization, or further human review. Care is needed to avoid reinforcing dataset-specific biases or over-interpreting uncertain or evolving user behavior.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): 1255776/COE12.

References

[1] Davide Chicco and Giuseppe Jurman. 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* 16, 1 (2023), 4.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the ACL: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[3] Bilal Ghanem, Simone Paolo Ponzetto, and Paolo Rosso. 2020. Factweet: profiling fake news twitter accounts. In *Statistical Language and Speech Processing: 8th International Conference, Cardiff, UK, Proceedings 8*. Springer, 35–45.

[4] Anastasia Giachanou, Bilal Ghanem, Esteban A. Ríssola, Paolo Rosso, Fabio Crestani, and Daniel Oberski. 2022. The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data & Knowledge Engineering* 138 (2022), 101960.

[5] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier.

[6] Tarek Hamdi, Hamda Slimi, Ibrahim Bounhas, and Yahya Slimani. 2020. A hybrid approach for fake news detection in twitter based on user features and graph embedding. In *Distributed Computing and Internet Technology: 16th International Conference, Bhubaneswar, India, Proceedings 16*. Springer, 266–280.

[7] Guoyong Hu, Ye Ding, Shuhan Qi, Xuan Wang, and Qing Liao. 2019. Multi-depth graph convolutional networks for fake news detection. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*. Springer, 698–710.

[8] Jisu Kim, Jihwan Aum, SangEun Lee, Yeonju Jang, Eunil Park, and Daejin Choi. 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics* 64 (2021), 101688.

[9] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907* (2016).

[10] Mohammad Mahyoob, Jeehaan Al-Garaady, and Musaad Alrahaili. 2020. Linguistic-based detection of fake news in social media. *Forthcoming, International Journal of English Linguistics* 11, 1 (2020).

[11] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.

[12] Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada* 2 (2013), 234.

[13] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

[14] Joan Plepi, Flora Sakketou, Henri-Jacques Geiss, and Lucie Flek. 2022. Temporal Graph Analysis of Misinformation Spreaders in Social Media. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*. ACL, Gyeongju, Republic of Korea, 89–104.

[15] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, Muhammad Sabih, and Hocine Cherifi. 2021. Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter. *IEEE Access* 9 (2021), 139636–139656.

[16] Bhavtosh Rath, Aadesh Salecha, and Jaideep Srivastava. 2020. Detecting Fake News Spreaders in Social Networks using Inductive Representation Learning. In *2020 IEEE/ACM International Conference on ASONAM*. 182–189.

[17] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 127 (Nov. 2019).

[18] Antonela Tommasel, Juan Manuel Rodriguez, and Filippo Menczer. 2022. Following the Trail of Fake News Spreaders in Social Media: A Deep Learning Model. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22 Adjunct)*. ACM, NY, USA, 29–34.

[19] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* 33 (2020), 5776–5788.

[20] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv:1905.11881* (2019).