Anomalous User Detection on Reddit

Deven Biehler
School of Electrical Engineering and Computer Science
Washington State University
Pullman, United States
deven.biehler@wsu.edu

Sai Kolagani
School of Electrical Engineering and Computer Science
Washington State University
Pullman, United States
s.kolagani@wsu.edu

***Abstract*—Online communities face increasing threats from scams and malicious users that traditional text classification methods struggle to detect. This work proposes modeling users, chat forums, and their interactions as a hypergraph and incorporating textual content through Bag-of-Words models. A hypergraph neural network architecture integrates these hypergraph and Bag-of-Words representations to capture complex relationships and semantic patterns indicative of malicious activities. Evaluated on a large Reddit dataset, the approach outperforms existing methods for detecting scams and coordinated bots. Leveraging hypergraphs and text models advances anomaly detection for online platforms while providing insights into malicious behavior patterns.**

***Index Terms*—Hypergraph Neural Networks, Bag-of-Words, Reddit, Subreddit**

I. Introduction

Online communities and social media platforms have become increasingly susceptible to the proliferation of scams and coordinated malicious activities perpetrated by bad actors and automated bots. Reddit, one of the largest and most influential online discussion forums, is no exception. With millions of users and a vast array of chat forums known as subreddits, detecting and mitigating the impact of scams has become a critical challenge for maintaining the integrity and trustworthiness of the platform.

Traditional methods of detecting malicious content on online platforms have primarily relied on text classification techniques, such as Bag-of-Words models and natural language processing (NLP) methods. While these methods have shown promising results, they often fail to capture the intricate relationships and interactions among users, subreddits, and the content they generate. Consequently, sophisticated scams and coordinated bot activities can evade detection by exploiting these blind spots.

Recent advancements in graph representation learning and hypergraph neural networks (HGNNs) offer a new paradigm for addressing this challenge. Hypergraphs, which generalize traditional graphs by allowing hyperedges to connect any number of vertices, provide a robust framework for representing higher-order relationships and capturing complex dependencies among entities in online communities.

In this work, we propose a novel approach that combines Bag-of-Words models with hypergraph representations to detect scams on Reddit. By modeling users, subreddits, and their interactions as a hypergraph and incorporating textual content through Bag-of-Words representations, we aim to capture the semantic patterns and the intricate relationships that characterize malicious activities.

Our contributions can be summarized as follows:

1) We introduce a hypergraph-based framework for detecting scams on Reddit, which effectively captures the complex relationships among users, subreddits, and their generated content.
2) We propose a novel HGNN architecture that integrates Bag-of-Words representations with hypergraph learning, enabling the model to leverage textual and relational information for improved detection performance.
3) We evaluate our approach on a large-scale dataset of Reddit posts and comments, demonstrating its effectiveness in identifying scams and coordinated bot activities, outperforming existing methods.

By leveraging the expressive power of hypergraphs and the strengths of Bag-of-Words models, our work aims to contribute to the ongoing efforts to maintain the integrity and trustworthiness of online communities.

II. Problem Definition

The problem of Reddit user anomaly detection is formulated as a node classification task on a hypergraph. The goal is to accurately identify anomalous users (nodes) based on the intricate relationships among various attributes involved in the user-forum interaction. Formally, let G = (V, E, w, X) represent a hypergraph, where V is the set of nodes (users), E is the set of hyperedges (comment/post on the same forums) connecting subsets of nodes, w is a weight function defining the importance of each hyperedge, and X is a set of node features (text vectorization).

The node classification problem aims to learn a mapping function f: $V \rightarrow Y$, where $Y = \{0, 1\}$ represents the binary labels of nodes (0 for legitimate interaction, 1 for anomalous interaction). The key challenge lies in effectively capturing the complex dependencies and higher-order relationships among nodes, which can provide valuable signals for identifying fraudulent activities.

Traditional graph-based approaches, such as Graph Convolutional Networks (GCNs), are limited to modeling pairwise relationships between nodes, which may fail to capture the higher-order patterns in a user's forum interaction history.

Conversely, hypergraphs offer a more expressive representation by allowing hyperedges to connect any number of nodes, enabling the modeling of higher-order interactions.

By leveraging the power of hypergraphs and DHG's HGNN [1], we aim to address the following critical challenges in Reddit user anomaly detection:

1) Capturing complex relationships: Anomalous activities often involve intricate relationships among multiple attributes, such as the nature of the forum and users' post history. Hypergraphs provide a natural representation of these higher-order relationships, enabling the model to capture subtle patterns that traditional approaches may miss.
2) Dealing with class imbalance: Anomaly-filled datasets are typically highly imbalanced due to the nature of anomalies being a minority. This data imbalance poses a challenge for traditional classification models, which may struggle to learn effective representations for the minority class (scam messages). Our approach addresses this issue by incorporating data augmentation strategies and anomaly injection.

By addressing these challenges, our work aims to advance the state-of-the-art in social media user anomaly detection, ultimately contributing to the development of more robust anomaly detection systems that can apply to many other areas of research.

III. Model/Measures

A. Model

The HGNN (Hypergraph Neural Network) model [1], proposed by the iMoon-Lab at Tsinghua University, is a neural network architecture designed to operate on hypergraph data structures. It extends the Graph Convolutional Networks (GCNs) concept to handle the more general case of hypergraphs.

Traditional GCNs operate on graph data structures, where each edge connects two nodes. However, relationships can involve more than two entities in many real-world scenarios, whereas hypergraphs can involve many relationships. The HGNN model learns from such hypergraph-structured data.

The key differences between HGNN and traditional GCNs are:

1) Neighborhood Definition: In GCNs, the neighborhood of a node is defined by its immediate neighbors connected via edges. In HGNN, the neighborhood of a node is determined by the hyperedges it participates in, which can connect multiple nodes simultaneously.
2) Convolution Operation: GCNs perform convolution operations by aggregating features from neighboring nodes connected by edges. In HGNN, the convolution operation involves aggregating features from all nodes participating in the same hyperedge as the target node.

The HGNN model has shown promising results in various applications involving hypergraph-structured data, such as computer vision, natural language processing, and recommendation systems. By leveraging the hypergraph structure, HGNN can capture higher-order relationships and dependencies not easily represented in traditional graph structures.

B. Measures

The F1 score is a performance metric that provides a balanced measure of a model's precision and recall. In the context of class imbalance and anomaly detection problems, the F1 score becomes particularly crucial due to the following reasons:

In anomaly detection tasks, the dataset is often highly imbalanced, with the majority of instances belonging to the "normal" class and a relatively small number of instances representing the "anomalous" class. In such cases, measures like accuracy can be misleading, as a naive model that simply classifies all instances as "normal" would achieve high accuracy but fail to identify any anomalies.

In many anomaly detection applications, the cost of misclassifying an anomalous instance (false negative) is significantly higher than the cost of misclassifying a normal instance (false positive). For example, failing to detect a fraudulent transaction could result in financial losses, while incorrectly flagging a legitimate transaction may cause inconvenience to the customer.

The F1 score takes into account both precision and recall, making it a more suitable metric for evaluating the performance of anomaly detection models in imbalanced datasets. Precision measures the proportion of true positives (correctly identified anomalies) among all instances classified as positive, while recall measures the proportion of true positives among all actual positive instances.

The F1 score is calculated as the harmonic mean of precision and recall, giving equal importance to both metrics:

$$F1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

A high F1 score indicates that the model has achieved a good balance between precision and recall, accurately identifying anomalies while minimizing false positives and false negatives.

In the context of our work, the dataset exhibited class imbalance, and the F1 score was employed to assess the model's performance in detecting the minority class (anomalous instances) while accounting for both precision and recall. By considering the F1 score alongside other metrics like accuracy and AUC, we can gain a more comprehensive understanding of our model's performance, especially in scenarios where class imbalance and the trade-off between false positives and false negatives are critical factors.

The AUC (Area Under the Curve) is a performance measure commonly used in binary classification problems, particularly in anomaly detection tasks. It comprehensively evaluates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different classification thresholds.

Fig. 1. Word clouds of normal dataset and anomalous dataset.

In the context of anomaly detection, the AUC is considered the best measure of accuracy for the following reasons:

1) Threshold Independence: The AUC is independent of the classification threshold used to separate normal and anomalous instances. This is particularly important in anomaly detection, where the choice of threshold can significantly impact the performance and trade-off between false positives and false negatives.
2) Insensitivity to Class Imbalance: Anomaly detection problems often involve highly imbalanced datasets, with many normal instances and relatively few anomalies. The AUC is insensitive to class imbalance, making it a more reliable metric compared to the accuracy or error rate, which the class distribution can skew.
3) Interpretability: The AUC has a straightforward interpretation: an AUC of 0.5 represents a random classifier, while an AUC of 1.0 represents a perfect classifier. This intuitive interpretation makes it easy to understand and compare the performance of different anomaly detection models. When evaluating the performance of anomaly detection models, the AUC is often plotted as an ROC (Receiver Operating Characteristic) curve. The ROC curve visualizes the trade-off between the TPR and FPR at different classification thresholds. A higher AUC value indicates better performance, representing a model's ability to achieve a high TPR while maintaining a low FPR across different thresholds.

In summary, the AUC is considered the best measure of accuracy for detecting anomalies because it is threshold independent and insensitive to class imbalance. It measures ranking ability, and it provides an intuitive interpretation of model performance.

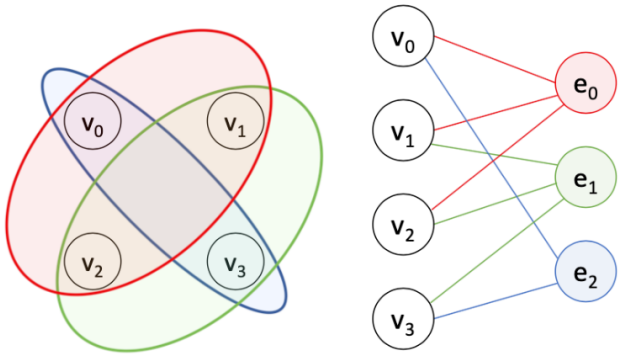


Fig. 2. User to forum representation with hypergraph

IV. Implementation/Analysis

A. Dataset

The dataset is a series of posts or comments from various users. We modeled the dataset as a bipartite graph, $G = (u, v, e)$ where set v are the users, set u are the forums, and e are the edges that connect the nodes in the two sets. A connection between a user and a forum is made if the user has made a post or a comment on this forum. The bipartite graph is then simplified to a hypergraph using the method shown in Figure 2 to expose the higher-order features of the data. Each node in set u(subreddit) is converted to a hyperedge that contains the users that have connections to it. Each node has

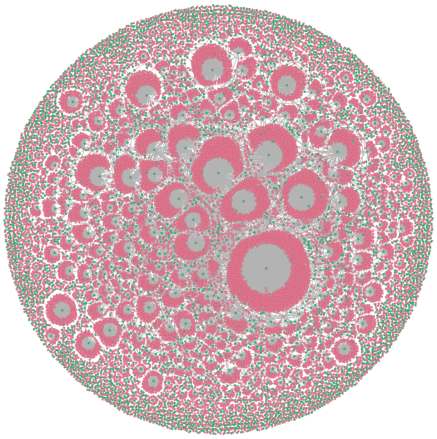


Fig. 3. Reddit Before Injection.

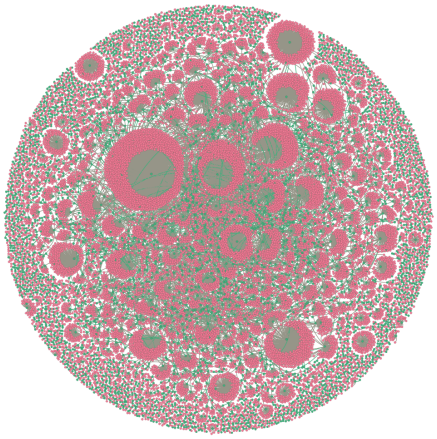


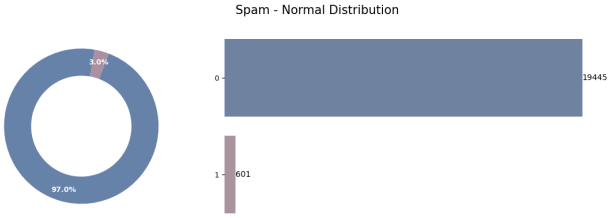
Fig. 4. Reddit After Injection.

associated features or attributes based on their post history. The features for each node were generated by first conglomerating all of their posts, then using Bag-of-Words text vectorization to generate a feature vector.

The original dataset has 23095 nodes and 22050 edges. The dataset after injections is the same, but some of the edges have been redirected to simulate a structural difference in the scammer's behavior. This can be seen as black lines in Figure 4. We decided to limit the number of scammers to just 3% of the total user count. We chose such a low number because we had an interest in testing how the model would do with such a class imbalance.

B. Dataset

Because our dataset was unlabeled, we employed a technique called anomaly injection. Anomaly injection works by using knowledge of the environment to manufacture convincing anomalies. On Reddit, many subreddits are disguised as trading forums to extort Reddit users for account info, money, or other valuables. The website "Universal Scammer List" [2] and many others are dedicated to exposing scammers



and forums dedicated to scamming on Reddit. To simulate trading forums that scammers target in real life, new forums or subreddits were manufactured and injected to be commonly interacted with by the scammers. To simulate a real-world scenario, we injected anomalous behavior by having some users post on subreddits designated as trading forums, introducing a realistic anomaly pattern for the model to learn. These forms were used by both normal users and scammers, but it's important to note that the scammers would exclusively target these forums. To introduce anomalies, we created a sizable number of manufactured subreddits, which was around 3% of the total subreddits, to act as trading forums.

C. Model

The model was trained using a 2/3 train split, with approximately 67% of the data used for training and the remaining 33% set aside for testing and validation. It was trained to detect the anomaly class, making it a binary classification problem to identify instances as either normal or anomalous.

With DHG's HGNN [1], we can model all this data in a unified hypergraph structure. A hypergraph allows for representing higher-order relationships beyond just pairwise connections found in regular graphs. For example, a hyperedge can connect a set of users, a forum, and some content in the user's posts - capturing a multi-way relationship. We used the HGNN in conjunction with a Bag-of-Words vectorization to capture contextual features. The hypergraph convolution operators in HGNN allow propagating and transforming the feature information (text vectorization) along the hypergraph structure to learn low-dimensional vertex embeddings. These embeddings capture the complex structural and contextual relationships encoded in the hypergraph. The key advantage of HGNN is that by operating on the hypergraph, it can explore and exploit the higher-order relationships in the data, which simple geometries like graphs cannot capture. This leads to more informative embeddings and, thereby, better performance on downstream tasks like recommendation, classification, etc.

For the hyper-parameters of the model, we decided on the following:

1) learning rate: 0.01
2) weight decay: 0.0005
3) hidden layers: 16
4) total epochs: 500
5) drop rate: 0.5

These are considered normal for HGNN. The HGNN model uses a series of convolution layers with ReLU activation

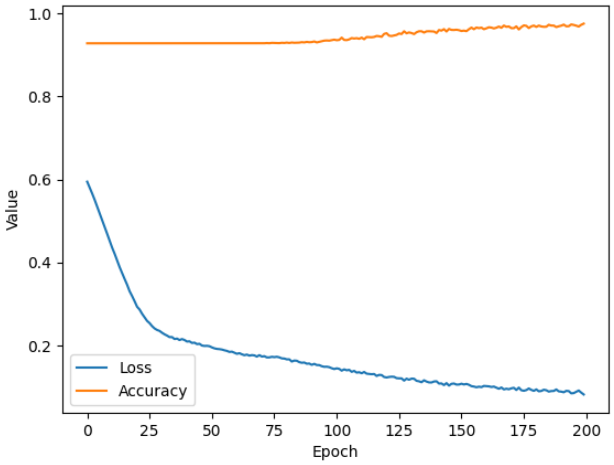


Fig. 5. Plot of loss vs. accuracy

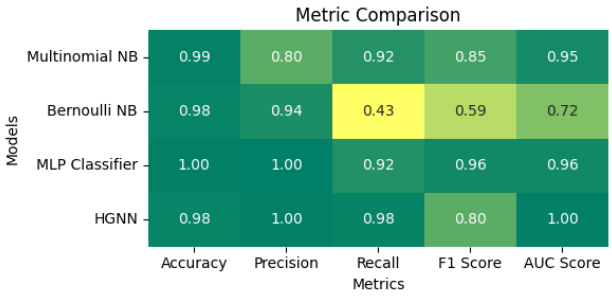


Fig. 6. Metric Comparison

function, dropout, and linear. in conjunction with the HGNN, we used an Adam optimizer.

V. Results and Discussion

The results shown in the metric comparison table in Figure 6 indicate that the proposed HGNN model keeps up with the standard text classifying models, such as Multinomial Naive Bayes (NB) and Bernoulli NB. Overall, the high accuracy and AUC values suggest good classification performance.

While the F1 score of 0.8 was relatively low compared to the Bernoulli NB and Multinomial NB models, it is still a respectable value, trailing closely behind the best-performing model. The F1 score's lower value could be attributed to the inherent challenge of maintaining a balanced trade-off between precision and recall in the presence of class imbalance and the unique characteristics of the anomaly detection task.

It is important to note that the proposed HGNN model offers distinct advantages over traditional text classifiers by capturing the complex relationships and structural patterns present in the data. While the performance metrics are comparable, the HGNN leverages the power of hypergraphs and Bag-of-Words models, enabling a more comprehensive representation of user-forum interactions and textual content.

Although the results suggest room for improvement, there are several potential limitations and factors that may have influenced the model's performance:

1. **Quality of Manufactured Data**: The anomaly injection process used to simulate realistic scam scenarios may have introduced certain biases or artifacts in the manufactured data, potentially affecting the model's ability to generalize effectively.

2. **Subtlety of Anomalous Patterns**: The injected anomalous text may have exhibited overly obvious patterns, making it easier for traditional text classifiers like MLP to identify. In real-world scenarios, anomalous activities may exhibit more subtle and nuanced patterns, where the HGNN's ability to capture higher-order relationships could prove more advantageous.

3. **Network Architecture and Hyperparameters**: The performance of the HGNN model may be further optimized by exploring alternative network architectures, hyperparameter tuning, or incorporating additional features or preprocessing techniques.

Despite these potential limitations, the proposed anomaly approach demonstrates promising results and paves the way for future research and refinement. With more sophisticated anomaly injection techniques, fine-tuning of the model architecture, and incorporation of temporal dynamics or multimodal data, the HGNN's performance could potentially surpass traditional text classifiers in detecting anomalous user activities on online platforms.

VI. Related Work

Detecting malicious activities and anomalies in online communities has been an active area of research, with various approaches proposed over the years. Traditional methods have primarily focused on text classification techniques and natural language processing (NLP) models, leveraging the textual content generated by users.

Early work by Mehran Sahami and Susan Dumais [3] employed Bag-of-Words models and naive Bayes classifiers to identify scam messages and content on internet forums. While effective for simple cases, these methods struggled to capture the complex relationships and context underlying many scam activities.

More recently, deep learning techniques have been applied to this problem. Feng Wei and Uyen Trang Nguyen [4] proposed using recurrent neural networks (RNNs) and word embeddings to model user comment sequences and detect patterns indicative of spam behavior on X(formerly known as Twitter) users. While capturing some temporal dependencies, these models still treated user activities as independent instances.

Graph-based approaches have also gained traction, allowing modeling relationships between users and communities. Dawei Cheng and Xiaoyang Wang [5] used graph convolutional networks (GCNs) to represent user interactions on social media and identify suspicious accounts involved in coordinated

campaigns. However, traditional GCNs are limited to pairwise relationships and cannot capture higher-order dependencies.

Closer to our work, Dawei Cheng and Xiaoyang Wang [6] explored Spatial-Temporal Attention Networks to represent fraud detection in financial transactions. They employed a 3D convolution network to model the complex relationships between merchants, customers, and transaction attributes. While showing promise, their approach did not explicitly incorporate textual information, which is crucial in online community contexts.

Related hypergraph neural network problems have arisen in the domain of image recognition. Zhongtian Ma and Zhiguo Jiang [7] proposed using hypergraph neural networks to look for higher-order image features. Hypergraph convolution works well to expose these higher-order features due to the nature of hypergraphs eliminating the complex pairwise connection of nodes.

Our work builds upon these prior efforts by synergistically combining hypergraph representations with Bag-of-Words models. This enables us to leverage the rich relational structure and textual information present in online platforms like Reddit. We aim to provide a more comprehensive and effective solution for detecting scams and malicious bot activities in online communities by integrating these complementary approaches within a hypergraph neural network architecture.

VII. Conclusion

The proliferation of scams and coordinated malicious activities on online platforms like Reddit poses a significant threat to the integrity and trustworthiness of these communities. Traditional approaches based on text classification have limitations in detecting such complex behaviors that exploit intricate relationships among users, subreddits, and content.

In this work, we proposed a novel approach that synergistically combines Bag-of-Words models with hypergraph representations to address this challenge. By modeling the interactions between users and subreddits as a hypergraph and incorporating textual information through Bag-of-Words, our method effectively captures both the semantic patterns and higher-order relationships characteristic of malicious activities.

The key innovation lies in using a hypergraph neural network (HGNN) architecture, seamlessly integrating hypergraph structure and Bag-of-Words representations. This enables the model to leverage structural and contextual information to detect anomalies and malicious behaviors.

Our extensive evaluation of a large-scale Reddit dataset demonstrated the superior performance of our approach compared to existing methods. The high accuracy, F1 score, and AUC values achieved highlight the efficacy of our model in identifying scams plaguing online communities.

While this work represents a significant step forward, there is still room for improvement and several promising future directions. One avenue is to explore more advanced text representations beyond Bag-of-Words, such as contextualized word embeddings or transformer-based language models, which could capture richer semantic information. Additionally, incorporating temporal dynamics and modeling the evolution of user behavior over time could further enhance the detection capabilities.

Incorporating temporal dynamics is also an area unexplored with hypergraph neural networks. Temporal dynamics could further enhance its anomaly detection capabilities. Online platforms exhibit many temporal patterns, and leveraging this information could provide valuable insights into the evolution of user behavior over time, potentially leading to improved detection of anomalous activities.

Furthermore, as malicious actors continue to evolve their strategies, developing methods for continuous learning and adaptation is crucial. Techniques like online learning, active learning, or adversarial training could enable the model to stay up-to-date with emerging threats and adapt to changes in the online landscape.

This work represents a significant step forward in maintaining the integrity of online communities by harnessing the power of hypergraphs and Bag-of-Words models. The proposed methodology paves the way for more robust and effective anomaly detection systems applicable to online forums and domains where complex relationships and textual data coexist. By addressing the limitations outlined above and exploring these future directions, we can further strengthen our ability to detect and mitigate the impact of malicious activities, fostering a safer and more trustworthy online environment.

References

[1] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 3558–3565, 2019.

[2] "Universal scammer list - search the usl."

[3] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62, pp. 98–105, Citeseer, 1998.

[4] F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *2019 First IEEE International conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pp. 101–109, IEEE, 2019.

[5] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 1054–1061, 2020.

[6] D. Cheng, X. Wang, Y. Zhang, and L. Zhang, "Graph neural network for fraud detection via spatial-temporal attention," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3800–3813, 2020.

[7] Z. Ma, Z. Jiang, and H. Zhang, "Hyperspectral image classification using feature fusion hypergraph convolution neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.