





GLORIA: A Graph Convolutional Network-Based Approach for Review Spam Detection

Giuseppina Andresini^{1,2} , Annalisa Appice^{1,2} , Roberto Gasbarro¹,
and Donato Malerba^{1,2} 

¹ Department of Computer Science, University of Bari “Aldo Moro”, Bari, Italy
{giuseppina.andresini, annalisa.appice, donato.malerba}@uniba.it,
r.gasbarro1@studenti.uniba.it

² CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Bari, Italy

Abstract. Spam reviews contain untruthful content created with malevolent intent, to affect the overall reputation of a product, service or company. This content is commonly made by malicious users or automated programs (i.e., bots) that mimic human behaviour. With the recent boom of online review systems, performing accurate review spam detection has become of primary importance for a review platform, to mitigate the effect of malicious users responsible for untruthful content. In this work, we propose a review spam classification approach, named GLORIA, that adopts a graph representation of review data and trains a graph convolutional neural network for edge classification as a review spam detection model. In particular, GLORIA represents both users (i.e., authors of reviews) and products (i.e., reviewed items) as nodes of a heterogeneous graph, while it represents reviews as graph edges that connect each author of a review to the reviewed item. Features of users, products and reviews are associated with nodes and edges, respectively.

Experiments performed on publicly available review datasets prove the effectiveness of the proposed approach compared with some state-of-the-art approaches.

Keywords: Review Spam Detection · Graph Convolutional Networks · Heterogeneous Graph Learning · Edge Classification

1 Introduction

With the continuous development of technology and the ubiquitous presence of network-based services in our everyday life, it has become very common to make online purchases of products and services. With the rapid spread of e-commerce services, user reviews have become one of the most influential factors in purchase decisions of customers [14]. Consequently, e-commerce marketplaces are nowadays the most important target of spammers, which have the malicious goal of manipulating the reputation of products and brands, to either promote

or criticize products and services. Positive and negative opinions can greatly influence a company’s business. For this reason, review spam detection is a crucial problem to address for guaranteeing the reliability of products and services.

Spam reviews are described as untruthful or deceptive opinions that are posted on online commerce platforms in an attempt to manipulate the public perception (in a positive or negative manner) of specific products or services presented on the affected platforms [11]. In the last years, automatic review spam detection has attracted the attention of machine learning, natural language processing and deep learning researchers due to the difficulty of recognizing fake reviews by manually reading their content [9]. On the other hand, over a few years, professional spammers have greatly increased and improved their writing techniques, to evade detection tools that base review spam detection on the analysis of text content only. Most of the existing review spam detection approaches focus on extracting robust, engineered features from both review contents and reviewer behaviours [2, 7, 17, 23], but in the past decade, several approaches have been developed, to leverage the social interaction between users and enhance the feature space of review spam detection problems with contextual information.

In this paper, we perform a step forward in this “social” research direction. In fact, we investigate the use of a heterogeneous graph representation of review data, to capture the relationships between products, users and reviews and gain accuracy in problems of review spam detection. For this purpose, we propose a Graph Convolutional Network (GCN) approach, named **GLORIA** (**G**raph **co**Lu**t**io**n**al Network for **R**ev**I**ew **s**p**A**m), that learns spatial convolutions on the graph representation of review data, to take advantage of the expressive structural information enclosed in graphs. In particular, the proposed approach implements a heterogeneous bipartite graph used as input to a Crystal GCN [28]. This architecture has been proven effective in the context of chemical material property prediction [8]. Traditional GNN algorithms perform convolutions using a shared weight matrix for all neighbours of a node by neglecting the difference of interaction between neighbours. Instead the Crystal GNN first aggregates neighbour vectors and then performs convolutions on the aggregated neighbour vectors. To the best of our knowledge, this is the first study that explores the use of Crystal GCN in review spam detection problems by showing how the proposed approach can gain accuracy compared to shallow and deep neural models trained neglecting the graph structure of data. An issue of review spam detection problems is that spam data are highly skewed. The imbalance of malicious data is a common condition in several cybersecurity problems (e.g., malware [4], fraud [22] or intrusion [3] detection), as well as in remote sensing problems (e.g., [6]). In this study, we handle the imbalanced condition of review spam data by training the Crystal GCN model with the sigmoid focal loss. This choice bases on [20] that shows how the sigmoid focal loss can help a neural model in focusing on rare samples. In this study, we show that the sigmoid focal loss is better suited than the traditional cross-entropy loss, to handle the imbalance condition of review spam data. Finally, we analyse the topological structure of graphs by showing how the exploration of the centrality of products and users in

the graph representation of review data may disclose useful knowledge to explain characteristics of reviews and possible spam patterns.

The paper is organized as follows. Section 2 overviews the related work. Section 3 describes the proposed approach. Section 4 describes the benchmark data collections adopted in the experimental study, describes the experimental setting and discusses the relevant results. Finally, Sect. 5 draws conclusions and outlines the future directions of this work.

2 Related Work

The research in the field of review spam detection has received great attention in the last years. Several machine learning approaches have been recently designed to disentangle spam reviews from non-spam reviews [9, 16]. In particular, the seminal study of [17] started the investigation of the task of review spam detection in the context of product reviews.

Recent research trends have started exploring deep learning approaches in problems of review spam detection [2, 5, 7, 24, 31]. In [2, 5] a multi-view, deep learning approach is described for review spam detection. The proposed approach combines embeddings of textual features, extracted with Word2Vec and BERT models, and behavioural reviewer features to improve the accuracy of a review classifier trained through a multi-input, deep neural network. [7] describes a combination of Word2Vec and Convolutional Neural Networks (CNNs), to learn a document-level representation of reviews. Finally, a Bi-directional LSTM is used for review classification. The work in [24] adopts word embeddings trained on an Amazon review dataset using the Continuous Bag-of-Words (CBOW) algorithm. Finally, it trains a review classification model that combines CNNs and Gated Recurrent Neural Networks. A CNN is also trained in [31], to extract semantic information from the text of reviews by exploiting convolution and pooling operations.

Although all the above studies describe feature-based approaches that rely on an effective way to extract and learn features (from both reviews and reviewers), they ignore relationships between users, products and reviews. On the other hand, a few recent studies have started the investigation of the effectiveness of graphs as data modelling approaches of review spam data. The study of [27] first adopts a heterogeneous graph to represent reviewers, reviews and stores, through different categories of nodes. The review graph is used to infer the truthiness of reviews, honesty of reviewers and reliability of stores. [26] explores an unsupervised review spam detection approach that resorts to clustering, to identify communities of users with similar spam behaviours. [25] describes the use of a heterogeneous graph to connect users to reviews and analyses how graph meta-paths may help in recognizing review spam.

The recent studies that have adopted a graph representation of review data have also paved the way for leveraging GCNs in review spam detection problems. Although GCNs have recently gained great attention in several domains (e.g., recommendation systems [29] and chemical properties predictions [28]), a few

studies have explored GCNs in review spam detection problems [1, 19, 30]. [1] trains a GCN for node classification in spam bot detection problems. This study adopts a social graph representation of relationships between Twitter users (represented as nodes) and leverages both feature nodes and relationships between neighbour nodes for training a GCN that addresses spam bot detection as a problem of graph node classification. A social graph of Twitter user relationships is also adopted in [30] in combination with an Attention-based Graph Neural Network trained for spam bot detection. So both these studies consider a problem of graph node classification, and train a GCN to recognize spam bot communities. They label all messages produced by the member of bot communities in the spam class. Differently, our study accounts that a malicious user does not necessarily only produce review spam. Based upon this consideration, we focus on classifying reviews, instead of classifying reviewers.

Finally, [19] studies the review spam detection problem for Xianyu, that is one of the largest second-hand goods apps in China. In Xianyu, reviews are communication tools for buyers and sellers and the review action usually happens before purchases. As recognised by the authors of [19], this is different from the common use of reviews in other e-commerce systems, also considered in this study, where reviews are usually made by customers who have bought the products. Accounting for the peculiar characteristics of reviews in Xianyu, [19] adopts two graph representations of Xianyu reviews: a heterogeneous graph modelling relationships between users and review items and a homogeneous graph modelling similarities between review items. A review item denotes a review topic (e.g., “iPhone 6s”) that is associated in Xianyu with a sequence of review comments produced by (multiple) users on the specific topic. The nodes of the homogeneous graph are associated with the content features extracted from review items. [19] concatenates embeddings extracted through the GCN trained on two graphs to obtain the feature vector for the final classification of the review item. Differently, we consider the traditional e-commerce perspective with reviews written by users on products. So we use a single heterogeneous graph to represent reviews as relationships between users and products, and we associate review features to edges, while characteristics of users and products to nodes. Finally, we train a GCN model for graph edge classification, to classify each single review message.

3 The Proposed Method

In this Section, we describe the GLORIA approach. It adopts a graph representation of review data, where users and products are represented as heterogeneous graph nodes and the reviews as graph edges. Hence, GLORIA implements a GCN for edge classification in heterogeneous graphs.

Let us consider the input graph representation of review data as a heterogeneous bipartite graph defined as $\mathbf{G} = (U, P, R)$, where U , P and R correspond to the set of user nodes, product nodes and edges, respectively. Each edge $r = (i, j) \in R$ with $R \subseteq (U \times P) \cup (P \times U)$ defines the undirect relationship between a user node and a product node to express that the user reviewed the

product (and the product is reviewed by the reviewer). In addition, let us consider three mapping functions: $\phi_U: U \rightarrow \mathbf{X}^U$ that associates each node $u \in U$ to a feature vector in \mathbf{X}^U , $\phi_P: P \rightarrow \mathbf{X}^P$ that associates each node $p \in P$ to a feature vector in \mathbf{X}^P and $\phi_R: R \rightarrow \mathbf{X}^R$ that associates each edge $r \in R$ to a feature vector in \mathbf{X}^R . We process this graph representation of review data to train a GCN for edge classification.

The GCN takes both node feature vectors, edge feature vectors and adjacency matrix as input and passes them through a series of L layers. At each layer l , node embeddings are updated according with the Eq. 1 to create an intermediate hidden representation \mathbf{h}^l . In particular, at each \mathbf{h}^l , the GCN of GLORIA applies a crystal graph convolutional operator [28]. For each node $i \in U \cup P$, for each layer l , this operator learns a function $\mathbf{h}_i^l = \mathbf{h}_i^{l-1} + f(i)$ defined on the previous hidden $l - 1$ layer. $f()$ is formulated as follows:

$$f(i) = AGG_{(j \in \mathcal{N}(i))^k} (\sigma(\mathbf{z}_{(i,j)^k}^1 \mathbf{W}_\sigma^1 + \mathbf{b}_\sigma^1) \odot g(\mathbf{z}_{(i,j)^k}^1 \mathbf{W}_g^1 + \mathbf{b}_g^1)) \quad (1)$$

where $\mathcal{N}(i)$ is the set of neighbours of node i (i.e., $\mathcal{N}(i) = \{j | (i, j) \in R\}$) and $\mathbf{z}_{(i,j)^k}^1 = \mathbf{h}_i^l \oplus \mathbf{h}_j^l \oplus \phi_R(i, j)^k$ is the concatenation of embeddings computed at layer l for the feature vectors associated with node i , neighbour node $j \in \mathcal{N}(i)$ and the feature vector associated with the k -th edge between i and j . If $l = 0$ then embeddings return original feature vectors computed with ϕ_U and ϕ_P for user nodes and product nodes, respectively. \odot denotes the element-wise multiplication, \mathbf{W}_σ and \mathbf{W}_g denote the convolution weight matrix at layer l , while b_σ and b_g denote the bias at layer l , for both σ and g functions. In fact, each layer of the Crystal GCN applies both a sigmoid function (σ) [13] and a softplus function (g) [10]. Finally, the operator of aggregation (AGG) denotes the aggregation scheme used for grouping node embeddings generated by different edges relating multiple neighbours j to the same node i . In this study, we use the mean as the aggregation operator.

In particular, GLORIA comprises two graph convolutional layers (i.e., $l = 1, 2$). During the message-passing phase at layer l , the information of each node of the graph is updated based on the aggregation of the messages received from their immediate neighbours achieved in two hops. As such, each message-passing layer increases the receptive field of the GCN by one hop. As we perform two hops in GLORIA, we are able to model relationships between pairs of users, as well as relationships between pairs of products, in addition to the review relationships between users and products.

Figure 1 reports an example of the message-passing realized by GLORIA by considering the user node u_1 as target node. At $l = 0$, all neighbour nodes of u_1 are assigned to initial node feature vectors (by ϕ_U or ϕ_P). At $l = 1$, the information of both node features and edge features are concatenated and aggregated, while $\mathbf{h}_{p_1}^1$ and $\mathbf{h}_{p_2}^1$ are updated based on $f()$ (Eq. 1). In order to get the node embeddings available for target node u_1 at $l = 2$, embeddings of neighbour nodes $\mathbf{h}_{p_1}^1$ and $\mathbf{h}_{p_2}^1$ are concatenated and aggregated to update $\mathbf{h}_{u_2}^2$. Therefore, each node in the graph learns from all the neighbour nodes transitively achieved in two hops.

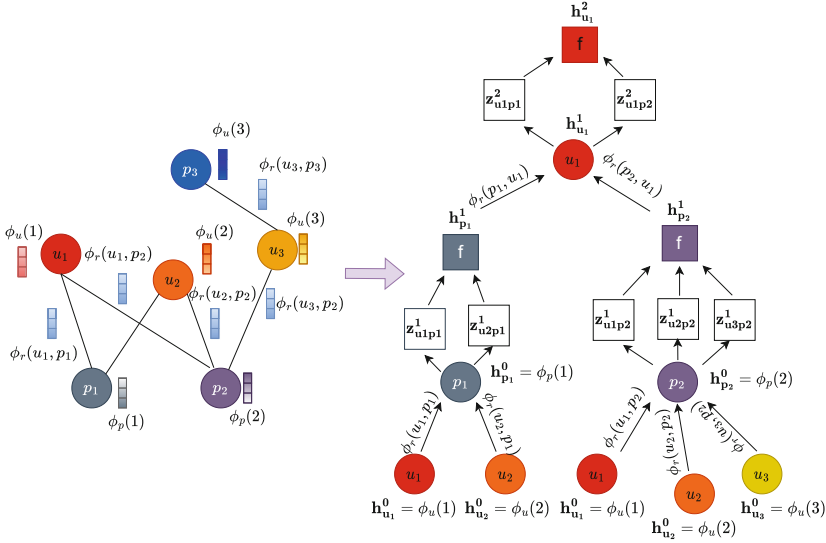


Fig. 1. Example of message-passing with user node u_1 as target in a Crystal GCN with two-hop neighbourhood

Finally, to deal with the expected imbalanced condition of review data in review spam detection problems, we use a sigmoid focal loss [20] for the final prediction:

$$\mathcal{SF} = - \sum_{k=1}^K \alpha(1 - \hat{y})^\gamma \log \hat{y} \quad (2)$$

where K corresponds to the number of classes in the dataset (i.e., spam, non-spam) and parameter γ adjusts the rate and reduces the loss for well-classified samples, to focus learning on hard misclassified samples. α is a weighting factor in range $(0, 1)$ to balance spam versus non-spam samples.

4 Experimental Setup

We performed experiments on two benchmark review datasets described in Sect. 4.1. The implementation details of GLORIA architecture, adopted in the experiments, are illustrated in Sect. 4.2. The experimental results are discussed in Sect. 4.3.

4.1 Data

We considered two datasets, namely Hotel and Restaurant, described in [23]. The two datasets contain reviews across 72 hotels and 129 restaurants, respectively, in the Chicago area. Each dataset contains reviews recorded by Yelp.com – a

Table 1. Summary of characteristics of Hotel and Restaurant datasets

| Dataset | #spam reviews | #non-spam reviews | #reviewers | #products |
|------------|---------------|-------------------|------------|-----------|
| Hotel | 779 | 5078 | 5123 | 72 |
| Restaurant | 8301 | 58716 | 16941 | 129 |

well-known large-scale online review site. In addition, both datasets were provided with ground-truth labels (spam and non-spam) in [23]. So they can be used for the evaluation of the accuracy of review spam detection approaches. Both datasets include information about products (e.g., category, price range, rating) and reviewers (e.g., number of friends, number of reviews), as well as plain-text reviews. Each review is associated with a reviewer and a product. In this study, we adopted the feature-vector representation of plain-text reviews described in [5]. In this study, each dataset was processed separately, as each domain has specific characteristics to take into account for the review spam analysis. A summary of the characteristics of both datasets is reported in Table 1. We note that the class distribution is imbalanced in both datasets with the “spam” minority class.

To perform the experimental study, we adopted the same split used in [5] with reviews sorted by the post date and the 80% of the oldest reviews selected for the training stage and the 20% of the newest reviews selected for the testing stage.

4.2 Implementation Details

We implemented GLORIA in Python 3¹. In particular, the GCN architecture was realized using PyTorch Geometric (PyG) 2.3, a geometric deep learning extension library for PyTorch. For each dataset, we conducted an automatic hyper-parameter optimization, using the tree-structured Parzen estimator algorithm, as implemented in the Hyperopt library. In particular, we selected the configuration of the hyper-parameters that achieved the highest F1 computed on the validation set extracted using 20% of the entire training according to the Pareto Principle, by considering spam as the positive class. The values of the search space of the hyper-parameters, automatically explored with the tree-structured Parzen estimator, are reported in Table 2.

The neural architecture of GLORIA comprises two Graph Convolutional Layers, a Dense layer and a Sigmoid layer [13] used for the final edge classification. The standard Rectified Linear Unit (ReLU) [12] was selected as the activation function for each hidden layer. A dropout layer was placed before each Graph Convolutional layer, to perform data regularisation and prevent overfitting. The neural network was trained with mini-batches by back-propagation, while the gradient-based optimization was performed using the Adam update rule [18]. The maximum number of epochs was set equal to 300. The early stopping app-

¹ <https://github.com/robertogasbarro/GLORIA>.

Table 2. Hyper-parameter search space for the multi-input neural network

| Hyper-parameter search-space values | |
|-------------------------------------|-------------------------------|
| Mini-batch size | $\{2^5, 2^6, 2^7, 2^8, 2^9\}$ |
| Learning rate | [0.0001, 0.001] |
| Dropout | [0, 1] |
| γ | [0, 1] |
| α | [1, 4] |

Table 3. F1 spam, F1 non-spam, Macro-F1 and AUC-ROC of GLORIA by using both BCE loss and SF loss for learning the GCN model. The best results are in bold.

| Dataset | Loss | F1 spam | F1 non-spam | Macro-F1 | AUC-ROC |
|------------|------|--------------|--------------|--------------|--------------|
| Hotel | BCE | 0.586 | 0.898 | 0.742 | 0.878 |
| | SF | 0.596 | 0.910 | 0.751 | 0.886 |
| Restaurant | BCE | 0.615 | 0.937 | 0.776 | 0.917 |
| | SF | 0.640 | 0.931 | 0.785 | 0.924 |

roach based on the lowest loss on the validation set was used, to obtain the best classification model.

4.3 Results and Discussion

We evaluate the performance of GLORIA to answer the following research questions:

- Q1 How does the accuracy of the proposed GCN-based approach change by varying the cost function?
- Q2 Does the defined GCN model gain accuracy compared to state-of-the-art review spam detection algorithms that neglect the graph structure of review data?
- Q3 Can the graph representation of review data disclose useful knowledge to explain the review domain better?

The accuracy performance of the analysed methods was measured in terms of F1 score computed on both the “spam” class and “non-spam” class, respectively (i.e., F1 spam and F1 non-spam), Macro-F1 (i.e., the average of F1 spam and F1 non-spam) and AUC-ROC. All these metrics were computed on the testing reviews of each dataset.

Sensitivity Analysis. We explored the sensitivity of the accuracy performance of GLORIA to the cost function adopted to learn the GCN model. To this aim, we compare the accuracy results obtained by using the binary cross-entropy (BCE)

Table 4. Competitor analysis: F1 spam, F1 non-spam, Macro-F1 and AUC-ROC of GLORIA, SVM and EUPHORIA. The best results are in bold.

| Dataset | Method | F1 spam | F1 non-spam | Macro-F1 | AUC-ROC |
|------------|----------|--------------|--------------|--------------|--------------|
| Hotel | SVM | 0.530 | 0.853 | 0.692 | 0.779 |
| | EUPHORIA | 0.592 | 0.887 | 0.740 | 0.813 |
| | GLORIA | 0.596 | 0.910 | 0.751 | 0.886 |
| Restaurant | SVM | 0.351 | 0.751 | 0.551 | 0.692 |
| | EUPHORIA | 0.372 | 0.781 | 0.576 | 0.706 |
| | GLORIA | 0.640 | 0.931 | 0.785 | 0.924 |

(defined as $\mathcal{BCE} = \sum_{k=1}^K y \log \hat{y}$) and the sigmoid focal (SF) loss (defined in Eq. 2), to perform the training stage of GLORIA.

Table 3 reports the F1 (spam), F1 (non-spam), Macro-F1 and AUC-ROC measured on the testing data by using both SF and BCE as loss function in both datasets. The results show that the use of the SF loss can help the GCN model to gain accuracy in both datasets. These results confirm the ability of the SF loss to improve the accuracy performance of a deep neural model in the presence of data showing a strong imbalanced condition. In fact, we can observe that the use of SF loss increases almost all the accuracy metrics in both the experimented datasets. The only exception is observed in the F1 (non-spam) calculated in Restaurant dataset, where the BCE loss performs better than SF loss. This is an expected outcome since the BCE loss is a cost function that considers samples of the two classes to have equal weights. Thus, the neural model can be learned with the BCE loss to recognise the majority class better (e.g., non-spam review in this study).

Competitor Analysis. We compare the accuracy performance of GLORIA to that of two competitors: SVM that learns a Support Vector Machine classifier and EUPHORIA that learns a multi-input deep neural model for review spam detection. We consider the SVM as a classification algorithm for this comparison since it has been already adopted in multiple related studies on review spam detection (e.g., [15, 21, 23]). On the other hand, EUPHORIA is a recent method described in [5] for review spam detection. Both competitors ignore the graph structure of review data.

Table 4 reports the F1 spam, F1 non-spam, Macro-F1 and AUC-ROC, of SVM, EUPHORIA and GLORIA, respectively. The results show that the highest accuracy is achieved by GLORIA, with EUPHORIA as runner-up of this experiment in both datasets. These results contribute to showing the effectiveness of resorting to a graph representation of review data and leveraging the graph structure of data to learn relationships between users and products, to improve the ability of the classification model to predict accurately the review spam.

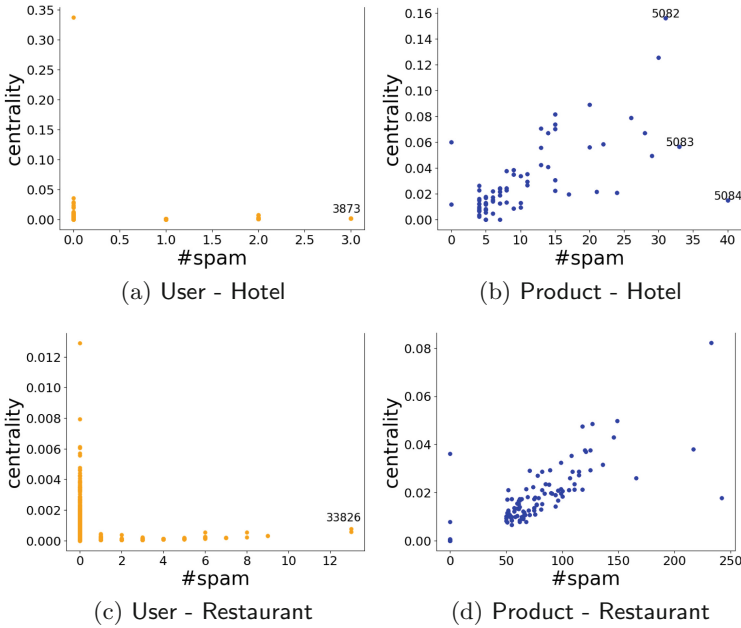


Fig. 2. Betweenness centrality (centrality, axis Y) with respect to the number of edges, which are labeled with the class “spam” on each node of the review graph (#spam axis X). Figures 2a and 2c refer to users, while Figs. 2b and 2d refer to products, in the review graphs of Hotel and Restaurant, respectively.

Qualitative Graph Analysis. Finally, we explore how the graph representation of the review data can disclose useful knowledge to explain the relationships between users, products and reviews in the spam class in the considered datasets. For this purpose, we analyse the betweenness centrality of users and products. The betweenness centrality of a node in a graph measures the amount of influence of the node on the flow of information in the graph. In particular, for a given node within a graph, the betweenness centrality of the node is computed as the number of the shortest paths in the graph, which connect any pair of nodes passing through the node under study, on the total number of the shortest paths which connect any pair of nodes in the graph. A node with a high value of betweenness centrality can be seen as a bridge that, if removed, could disrupt connections between other nodes in the graph.

Due to the main focus of this problem on the class “spam”, Fig. 2 shows the betweenness centrality (axis Y) plotted with respect to the number of edges labelled with the class “spam” on each graph node. As GLORIA adopts a heterogeneous graph, we show the betweenness centrality for both users (Figs. 2a and 2c) and products (Figs. 2b and 2d). These results show that the betweenness centrality of a node tends to increase as the number of spam reviews involving the node increases. This trend is more evident in the products than in the users,

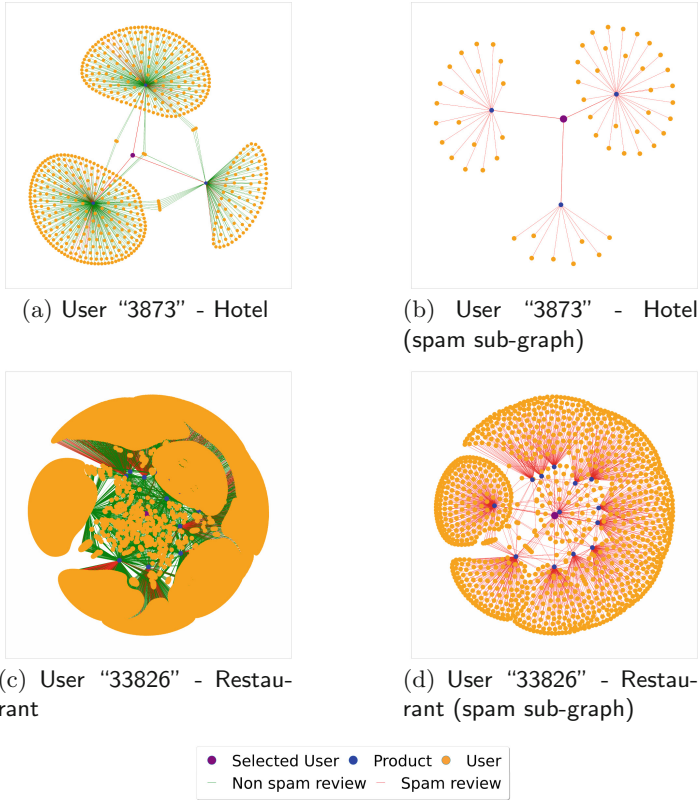


Fig. 3. Sub-graphs rooted in the users: “3873” of Hotel and “33826” of Restaurant. Both sub-graphs are produced with two hop levels. Figures 3b and 3d report the projection of sub-graphs shown in Figs. 3a and 3c on the edges labeled with the class “spam”.

since there is a large number of users who produced zero spam reviews, while a small number of products received zero spam reviews.

Figure 3 shows the sub-graphs rooted in the users: “3873” of Hotel and “33826” of Restaurant. These users are identified according to the plots reported in Figs. 2b for Hotel and 2d for Restaurant as the users who produced the highest number of reviews labeled in the class “spam” in the two datasets. The two sub-graphs are produced with two hop levels. In particular, Figs. 3a and 3c show the entire sub-graphs rooted in the selected users “3873” and “33826”, respectively. Figures 3b and 3d show the projection of these sub-graphs on the edges labeled in the class “spam”. The sub-graphs rooted in the users “3873” and “33826” show the products for which these two users are spammers into Hotel and Restaurant, respectively. In both cases, the sub-graphs highlight that “multiple” reviewers produced spam reviews on the same target products. This suggests that several spammer profiles co-operated to produce malicious spam

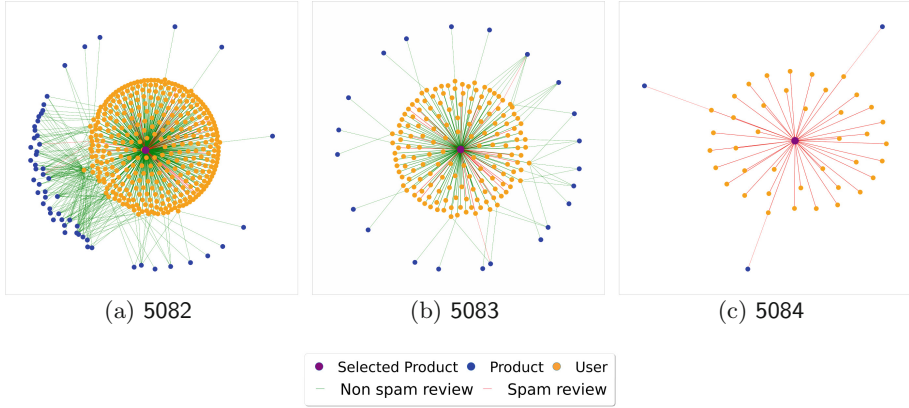


Fig. 4. Subgraphs of products “5082”, “5083” and “5084” of Hotel dataset. These products have achieved the higher number of spam as reported in Fig. 2b.

on the same target products. Hence, the activity of a spammer on a product may attract the attention of further spammers on the same target product.

Figure 4 shows the sub-graphs rooted in products “5082”, “5083” and “5084” of Hotel, which are the three products that received the highest number of spam reviews in Fig. 2b. We note that the product “5082” has the highest *betweenness centrality* in Fig. 4. Instead, the product “5084” has a low *betweenness centrality*, while the product “5083” has a medium *betweenness centrality* in Fig. 4. Consistently with this analysis of *betweenness centrality*, the node density is higher in the sub-graph rooted in product “5082” than in the sub-graphs rooted in product “5083” and product “5084”, respectively. In addition, the sub-graph rooted in product “5084” shows that all reviews produced on this product belong to the class “spam”. These malicious reviews were produced by users who created these single reviews (with the exception of two users who created two reviews and both these reviews were spam). This suggests that a possible malicious behaviour is observable in products with low *betweenness centrality*.

5 Conclusion

In this paper, we illustrate a GCN approach for review spam detection, which takes advantage of relationships between users, products and reviews by resorting to a graph-based representation of review data and training GCN model for edge classification of reviews. The experiments performed on two benchmark datasets prove the accuracy of the proposed approach compared with two baselines that are SVM and a multi-view deep learning-based approach, respectively. In addition, we show that the analysis of the *betweenness centrality* of products and users allows us to extract useful knowledge to explain review data by disclosing possible review spam patterns. As future work, we plan to continue the investigation of how knowledge explaining the review graph topology can be

used during the GCN training stage, to help the learned classification model gain accuracy in detecting review spam. In addition, we plan to investigate the use of the graph-based representation of review data in an online setting, to explore how changes occurring over time in the graph topology may help to keep high accuracy detecting review spam in real-time. Finally, we plan to extend our approach to perform a link prediction task to predict future behaviours of spammers.

Acknowledgments. The work of Giuseppina Andresini and Donato Malerba was supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU. The work of Annalisa Appice was partially supported by project SERICS (PE00000014) under the NRRP MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. The authors wish to thank Raffaele Scaringi for the helpful discussion on Graph Neural Networks.

CRediT Authorship Contribution Statement

Giuseppina Andresini. Conceptualization, Methodology, Data curation, Investigation, Validation, Supervision, Visualization, Writing - original draft, Writing - review & editing. **Annalisa Appice:** Conceptualization, Methodology, Investigation, Validation, Supervision, Writing - original draft, Writing - review & editing. **Roberto Gasbarro:** Methodology, Software, Investigation, Data curation, Visualization, Writing - review & editing. **Donato Malerba:** Conceptualization, Writing - review & editing.

References

1. Ali Alhosseini, S., Bin Tareaf, R., Najafi, P., Meinel, C.: Detect me if you can: spam bot detection using inductive representation learning. In: Companion Proceedings of the 2019 World Wide Web Conference, WWW 2019, pp. 148–153. Association for Computing Machinery, New York (2019). <https://doi.org/10.1145/3308560.3316504>
2. Andresini, G., Iovine, A., Gasbarro, R., Lomolino, M., de Gemmis, M., Appice, A.: Review spam detection using multi-view deep learning combining content and behavioral features. In: CEUR Workshop Proceedings, vol. 3340, pp. 87–98 (2022)
3. Andresini, G., Appice, A., Caforio, F.P., Malerba, D., Vessio, G.: ROULETTE: a neural attention multi-output model for explainable network intrusion detection. *Expert Syst. Appl.* 117144 (2022). <https://doi.org/10.1016/j.eswa.2022.117144>
4. Andresini, G., Appice, A., Malerba, D.: Dealing with class imbalance in android malware detection by cascading clustering and classification. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) *Complex Pattern Mining. SCI*, vol. 880, pp. 173–187. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-36617-9_11
5. Andresini, G., Iovine, A., Gasbarro, R., Lomolino, M., de Gemmis, M., Appice, A.: Euphoria: a neural multi-view approach to combine content and behavioral features in review spam detection. *J. Comput. Math. Data Sci.* **3**, 100036 (2022). <https://doi.org/10.1016/j.jcmds.2022.100036>

6. Appice, A., Malerba, D.: Segmentation-aided classification of hyperspectral data using spatial dependency of spectral bands. *ISPRS J. Photogramm. Remote. Sens.* **147**, 215–231 (2019). <https://doi.org/10.1016/j.isprsjprs.2018.11.023>
7. Bhuvaneshwari, P., Rao, A., Robinson, H.: Spam review detection using self attention based CNN and bi-directional LSTM. *Multimed. Tools Appl.* **80**, 1–18 (2021)
8. Cheng, J., Chunkai, Z., Dong, L.: A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater.* **2** (2021). <https://doi.org/10.1038/s43246-021-00194-3>
9. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N., Al Najada, H.: Survey of review spam detection using machine learning techniques. *J. Big Data* **2**(1), 1–24 (2015). <https://doi.org/10.1186/s40537-015-0029-9>
10. Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., Garcia, R.: Incorporating second-order functional knowledge for better option pricing. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS 2000*, pp. 451–457. MIT Press, Cambridge (2000)
11. Ferrara, E.: The history of digital spam. *Commun. ACM* **62**(8), 82–91 (2019). <https://doi.org/10.1145/3299768>
12. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: *AISTATS*, pp. 315–323. JMLR.org (2011)
13. Han, J., Moraga, C.: The influence of the sigmoid function parameters on the speed of backpropagation learning. In: Mira, J., Sandoval, F. (eds.) *IWANN 1995. LNCS*, vol. 930, pp. 195–201. Springer, Heidelberg (1995). https://doi.org/10.1007/3-540-59497-3_175
14. Heydari, A., ali Tavakoli, M., Salim, N., Heydari, Z.: Detection of review spam: a survey. *Expert Syst. Appl.* **42**(7), 3634–3642 (2015). <https://doi.org/10.1016/j.eswa.2014.12.029>
15. Hussain, N., Mirza, H., Hussain, I., Iqbal, F., Memon, I.: Spam review detection using the linguistic and spammer behavioral methods. *IEEE Access* **8**, 53801–53816 (2020). <https://doi.org/10.1109/ACCESS.2020.2979226>
16. Hussain, N., Turab Mirza, H., Rasool, G., Hussain, I., Kaleem, M.: Spam review detection techniques: a systematic literature review. *Appl. Sci.* **9**(5) (2019)
17. Jindal, N., Liu, B.: Opinion spam and analysis. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM 2008*, pp. 219–230. Association for Computing Machinery, New York (2008)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: *ICLR* (2014)
19. Li, A., Qin, Z., Liu, R., Yang, Y., Li, D.: Spam review detection with graph convolutional networks, pp. 2703–2711 (2019). <https://doi.org/10.1145/3357384.3357820>
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(2), 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
21. Lin, Y., Zhu, T., Wu, H., Zhang, J., Wang, X., Zhou, A.: Towards online anti-opinion spam: spotting fake reviews from the review sequence. In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 261–264 (2014). <https://doi.org/10.1109/ASONAM.2014.6921594>
22. Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.S., Zeineddine, H.: An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access* **7**, 93010–93022 (2019). <https://doi.org/10.1109/ACCESS.2019.2927266>

23. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: Kiciman, E., Ellison, N.B., Hogan, B., Resnick, P., Soboroff, I. (eds.) Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, 8–11 July 2013. The AAAI Press (2013). <https://doi.org/10.1609/icwsm.v7i1.14389>
24. Ren, Y., Ji, D.: Neural networks for deceptive opinion spam detection: an empirical study. *Inf. Sci.* **385–386**, 213–224 (2017). <https://doi.org/10.1016/j.ins.2017.01.015>
25. Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N.: NetSpam: a network-based spam detection framework for reviews in online social media. *IEEE Trans. Inf. Forensics Secur.* **12**, 1585–1595 (2017). <https://doi.org/10.1109/TIFS.2017.2675361>
26. Soliman, A., Girdzijauskas, S.: AdaGraph: adaptive graph-based algorithms for spam detection in social networks. In: El Abbadi, A., Garbinato, B. (eds.) NETYS 2017. LNCS, vol. 10299, pp. 338–354. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59647-1_25
27. Wang, G., Xie, S., Liu, B., Yu, P.S.: Identify online store review spammers via social review graph. *ACM Trans. Intell. Syst. Technol.* **3**(4) (2012). <https://doi.org/10.1145/2337542.2337546>
28. Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018). <https://doi.org/10.1103/PhysRevLett.120.145301>
29. Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., Leskovec, J.: Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, pp. 974–983. Association for Computing Machinery, New York (2018). <https://doi.org/10.1145/3219819.3219890>
30. Zhao, C., Xin, Y., Li, X., Zhu, H., Yang, Y., Chen, Y.: An attention-based graph neural network for spam bot detection in social networks. *Appl. Sci.* **10**(22) (2020). <https://doi.org/10.3390/app10228160>
31. Zhao, S., Xu, Z., Liu, L., Guo, M.: Towards accurate deceptive opinion spam detection based on word order-preserving CNN. *Math. Probl. Eng.* **2018** (2018). <https://doi.org/10.1155/2018/2410206>