A REVIEW ON FAKE PROFILE DETECTION USING DEEP LEARNING WITH BERT AND GCN MODELS

M.Anantha Lakshmi, K.Pavani, K.N.S.S.Supraja,V.Vasantha Lakshmi, K.Kalyani, A.V.S.S.R.Rao

Department of Computer Science and Engineering

Sasi Institute of Technology and Engineering, Tadepalligudem

Abstract

The rise of social media has led to an increase in fake profiles, which are often used for misinformation, fraud, cybercrimes, and malicious activities. Detecting such profiles is a challenging task due to their evolving behaviour and advanced strategies. This project aims to develop a fake profile detection system using deep learning techniques, specifically BERT (Bidirectional Encoder Representations from Transformers) for text-based feature extraction and GCN (Graph Convolutional Networks) for analysing network structures. By integrating natural language processing (NLP) and graph-based learning, this approach provides a more robust and scalable solution for detecting fake profiles. The proposed method effectively distinguishes between real and fake users, representing a significant improvement over traditional machine learning techniques. The TwiBot-20 dataset, a benchmark dataset containing real and fake social media profiles, is used for training and evaluation.

Keywords : Social media profiles, BERT, GCN, Model training , Metrics.

Introduction

In recent years, the rise of social media platforms and online services has brought about a significant challenge: the widespread of fake profiles.These profiles are created with the intent to deceive, manipulate, or exploit other users, whether for malicious purposes like fraud, harassment, or misinformation, or simply for spamming and promoting fake content. Detecting and preventing fake profiles is crucial for maintaining the integrity of online platforms, ensuring user trust, and protecting individuals and organizations from various types of online threats. Traditional methods for detecting fake profiles, such as rule-based systems, often fall short in terms of scalability, adaptability, and effectiveness due to the evolving nature of fake profiles.

Nowadays, many users use Social Media Platforms (SMPs) to connect with their friends and family. People spend a significant amount of time on sites like Facebook, Instagram, and Twitter updating themselves about the world. Social media platforms verify the authenticity of registered users. However, some users hide their identities, and these people threaten the security of other users' data.

However, this exponential growth and accessibility have also led to significant vulnerabilities, such as the rapid increase of fake profiles, which pose serious threats to the integrity, trustworthiness, and security of social media. Fake profiles are artificial accounts created with the intent to deceive. These may impersonate real individuals, represent fictitious entities, or present themselves as authentic users. The consequences of such fake profiles extend far beyond simple annoyances:

1. Online Deception: Fake profiles may be used to carry out scams, including phishing schemes, extortion, and financial fraud, thereby causing monetary loss and emotional distress to unsuspecting users.

2. Identity Theft: Many fake profiles rely on stolen personal data and photographs to impersonate genuine individuals, leading to reputational damage, mistrust, and legal complications.

3. Propagation of Misinformation: These accounts can amplify false narratives, fake news, and misleading information that distort public opinion and create chaos during critical times, such as elections or global crises.

4. Cyber-bullying and Harassment: Fake profiles are frequently used for abusive behaviour, stalking, and trolling, resulting in mental health challenges for victims.

5. Impact on Businesses: Fake profiles can compromise the effectiveness of marketing campaigns, erode brand trust, and spread malicious content targeting businesses.

Fake profile detection is a critical task in many domains, such as social media platforms, online marketplaces, and dating apps, where users may create fraudulent accounts to deceive others. Deep learning models, such as BERT and GCN are increasingly being used to tackle this problem. A fusion of these two models can improve the detection of fake profiles by leveraging the strengths of both.

By combining BERT and GCN for fake profile detection leverages the strengths of both textual analysis and relational data. BERT focuses on understanding the content of user profiles, detecting linguistic patterns indicative of fake text, while GCN analyses the social graph to spot suspicious relationships or unusual interactions that may point to fraudulent accounts. The fusion of these two models allows for a more robust, accurate, and generalizable system for detecting fake profiles in online platforms.

The primary goal of fake profile detection is to identify and flag fraudulent or deceptive profiles on online platforms, such as social media networks, online dating apps, e-commerce sites, or forums. Fake profiles may be created for malicious purposes such as spreading misinformation, scamming users, or creating a false impression to

gain trust or influence. Detecting these profiles is essential to maintain the integrity of online environments.

When using deep learning techniques, particularly BERT and GCN the objective is to leverage these models' strengths to accurately and efficiently identify fake profiles by analysing both textual and graph-based data. This approach is significant because it enhances platform safety, reduces fraudulent activities, improves user experience, and combats the spread of misinformation. By leveraging the strengths of both BERT for understanding text and GCN for analysing network structures, a fused deep learning model can detect subtle patterns that indicate fake profiles, offering a powerful tool to maintain the integrity of online platforms and protect users from malicious activity.

The Proposed Architecture

BERT (Bidirectional Encoder Representations from Transformers):

BERT is a pre-trained transformer model developed by Google, specifically designed to process and understand natural language. BERT works bi-directionally, meaning it understands the context of each word in a sentence by considering both its left and right context.

BERT is pre-trained on a large corpus of text, such as Wikipedia and books, using two key tasks:

- **Masked Language Modelling (MLM)**: In MLM, some words in a sentence are randomly masked, and the model must predict the missing words. This helps the model learn bidirectional context (i.e., understanding words based on both the preceding and following words).
- **Next Sentence Prediction (NSP)**: This task trains the model to predict if one sentence follows another, allowing BERT to capture sentence-level relationships and dependencies, making it useful for understanding the flow of text in a profile or conversation.

This pre-training allows BERT to learn how language works at a deep level, including grammar, syntax, and semantic meanings.

2. Fine - Tuning:

Once pre-trained, BERT can be fine-tuned for specific tasks like fake profile detection. This is done by training the model on a labelled dataset that contains real and fake profiles. Fine-tuning enables BERT to adapt its knowledge of language to recognize specific patterns of deception in profiles.

Fine-tuning process:

- **Input**: A user profile's textual data (such as profile description, posts, etc.) is tokenized and processed as input to BERT.
- **Textual Embedding**: BERT generates embedding for each word or token in the input profile. These embedding capture the contextual meaning of each word, considering the words around it.

2. BERT Pre-Training:

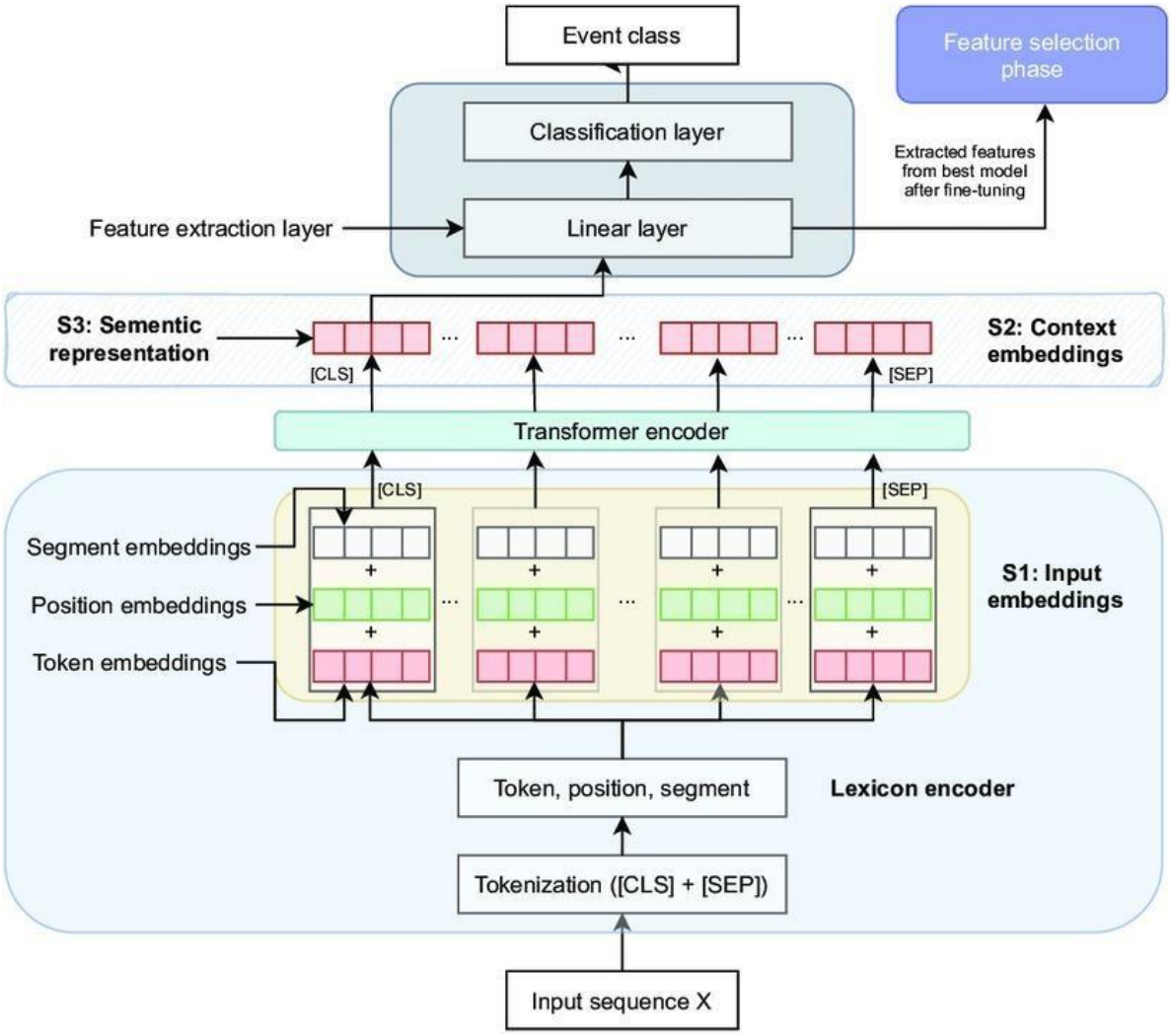

Fig 1 : BERT Training Model

- **Output**: After passing through several layers of the transformer network, BERT outputs a set of representations (embedding) that reflect the full context of the profile text.
- **Classification Layer**: At the final layer, a classifier (usually a feed-forward neural network) is applied to predict whether the profile is fake or real based on the embedding generated by BERT. The classifier typically outputs a probability score indicating the likelihood that the profile is fake.

GCN (Graph Convolutional Network):

GCN offer a powerful method for analyzing graph-based data, such as social networks, where relationships between users (i.e., interactions, friendships, followers, etc.) play a crucial role in identifying suspicious or fake accounts. Since fake profiles often exhibit abnormal patterns in how they interact with other users or how isolated they are within a social network. GCN is well-suited to detect these structural anomalies.

GCN (Graph Convolutional Network):

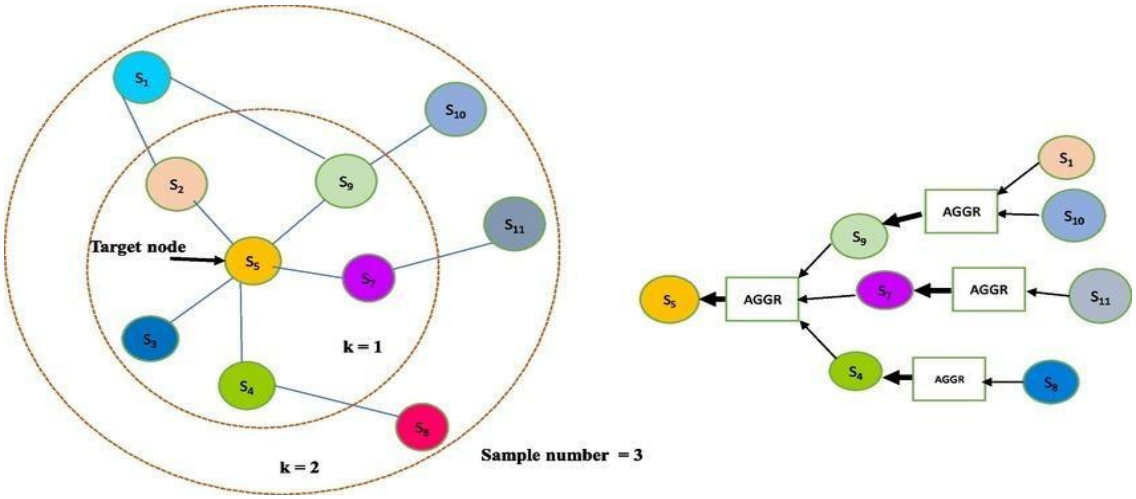

Fig 2 : GCN Architecture

GCNs extend traditional convolutional neural networks (CNNs) to graph data, allowing them to learn node representations (embedding) by aggregating information from neighboring nodes (users) in the network. In the context of fake profile detection, GCNs can help identify suspicious network behavior that could indicate a profile is fake .

1. Graph Representation of Social Networks

A social network (or user interaction network) can be represented as a graph, where:

- Nodes represent individual users.
- Edges represent relationships or interactions between users (e.g., friendships, followers, likes, comments, etc.
- Isolated nodes with few or no connections could represent fake accounts.
- Dense clusters or groups of interconnected users could indicate real, active users.
- Abnormal patterns in connectivity, such as excessive links to suspicious profiles or unusual network structures, might suggest fraud or manipulation.

2. Graph Convolution

- In GCN, graph convolution is performed to update each node's feature representation by aggregating the features of its neighboring nodes. This step is important because a user's characteristics can often be inferred not just from their individual profile but also from the profiles of their neighbors.
- A graph convolution layer takes the node features and the structure of the graph (adjacency matrix, which represents the connections between nodes) to aggregate information from neighboring nodes and update the node’s feature vector.

- For each user (node), the GCN aggregates information from neighbouring users. For example, if a user has connections with many suspicious or isolated profiles, the aggregated features may indicate abnormal behaviour or that the user is likely to be a fake profile.
- After multiple layers of graph convolution, each node has an updated embedding that represents both its individual features and its relationships with other users in the network. These final node embedding can capture complex patterns, such as being part of an isolated group or being linked to a disproportionate number of other fake profiles.

Fusing BERT and GCN for Fake Profile Detection

The fusion of BERT and GCN allows us to leverage the strengths of both models: BERT's ability to process and understand textual data and GCN's ability to analyse relationships and structural patterns within the network.

1. **Text Features with BERT**:

- First, textual data from user profiles (such as descriptions, posts, and comments) is processed by BERT to extract meaningful features. BERT generates embeddings that capture the semantic meaning of the text.
- These features can be used to classify whether a user's profile text is fake or real based on linguistic patterns.

2. **Graph Features with GCN**:

- Second, a graph is constructed where nodes represent users and edges represent interactions between users (e.g., friendships, messages).
- GCN is applied to this graph to learn the network structure and detect anomalies, such as isolated users or abnormal interaction patterns.

3. **Multimodal Fusion**:

- Once BERT and GCN process the respective data types (text and network structure), the results are fused. The fusion could be done by combining the embeddings from BERT with the node representations from GCN, followed by feeding them into a classifier (e.g., a neural network or SVM).
- This allows the model to make a final decision on whether a profile is fake by considering both the textual features and the relationship/network features together.

In summary, combining BERT and GCN for fake profile detection leverages the strengths of both textual analysis and relational data. BERT focuses on understanding the content of user profiles, detecting linguistic patterns indicative of fake text, while GCN analyses the social graph to spot suspicious relationships or unusual interactions that may point to fraudulent accounts. The fusion of these two models allows for a more robust, accurate, and generalizable system for detecting fake profiles in online platforms.

Dataset

The Twitbot-20 dataset contains information about Twitter accounts, both human and bot. It includes 37,438 entries, each representing a single Twitter user, with 20 features describing various aspects of their accounts.

Key Details:

- Type of Accounts**:** Labelled as either a human or a bot (account-type column: 0 = human, 1 = bot).
- Account Information**:** Includes data like creation date, number of followers, friends, and tweets.
- Profile Settings: Flags for default profile settings and images.
- Activity Metrics: Average tweets per day and total number of statuses.
- Verification: Indicates if the account is verified (verified column).
- Profile Details: Language, location, profile images, and description.

Results

Fake profiles can be created for malicious purposes such as spreading misinformation, phishing, or manipulating public opinion. So we have used BERT and GCN models for capturing the inherent connections in the social media data. BERT is employed to process and extract deep contextual embedding from the profile descriptions and messages. The GCN propagates and aggregates features over the graph, enabling the model to incorporate information from neighboring nodes. A classification layer (fully connected layer) is added after the GCN outputs to predict whether a profile is real or fake. We have used evaluation metrics such as accuracy, precision, f1-score and recall and Confusion matrix which indicates effective detection. The hybrid approach (combining BERT and GCN) leads to high accuracy and robustness, making it suitable for real-world deployment in social media platforms.

Conclusion

The fake profile detection system developed using deep learning techniques, specifically BERT and Graph Convolutional Networks (GCN), demonstrates a powerful and effective approach to identifying fraudulent accounts on social media platforms. By integrating BERT, which excels at understanding the contextual meaning of text data, with GCN, which captures the relational information between user accounts in a network, the system leverages both the individual behavior of users and their connections within the social graph. These embedding capture complex information about how genuine users communicate, compared to the often templated or inconsistent language used by fake profiles.

The integrated system exhibits a strong ability to distinguish between authentic and fake profiles, with minimal false positives and false negatives, making it a reliable solution for practical deployment. The application of BERT and GCN for fake profile detection presents a highly effective and comprehensive solution.

By combining advanced natural language processing with graph-based learning, the model not only enhances detection accuracy but also provides a scalable and adaptable framework for tackling the evolving challenges posed by fake profiles in online social networks. This method has the potential to significantly improve the integrity and trustworthiness of digital platforms, ensuring a safer and more authentic user experience.

References

[1] K. Umbrani, A. Jain, D. Shah, and A. Pile, "Fake Profile Detection using Machine Learning," in 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Pune, India, 2024, pp. 1-8. doi: 10.1109/ICETSIS61505.2024.10459570.

[2] Kudugunta, Sneha & Ferrara, Emilio. (2018). Deep Neural Networks for Bot Detection. Information Sciences. 467. 10.1016/j.ins.2018.08.019.

[3] Miller, Zachary & Dickinson, Brian & Deitrick, William & Hu, Wei & Wang, Alex. (2014). Twitter spammer detection using data stream clustering. Information Sciences. 260. 10.1016/j.ins.2013.11.016.

[4] Cresci, Stefano & Pietro, Roberto & Petrocchi, Marinella & Spognardi, Angelo & Tesconi, Maurizio. (2016).DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. IEEE Intelligent Systems. 31. 58-64. 10.1109/MIS.2016.29.

[5] Davis, Clayton & Varol, Onur & Ferrara, Emilio & Flammini, Alessandro & Menczer, Filippo. (2016). BotOrNot: A System to Evaluate Social Bots. 273-274. 10.1145/2872518.2889302.

[6] Miller, Zachary & Dickinson, Brian & Deitrick, William & Hu, Wei & Wang, Alex. (2014). Twitter spammer detection using data stream clustering. Information Sciences. 260. 10.1016/j.ins.2013.11.016.

[7] Cresci, Stefano & Pietro, Roberto & Petrocchi, Marinella & Spognardi, Angelo & Tesconi, Maurizio. (2016).DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. IEEE Intelligent Systems. 31. 58-64. 10.1109/MIS.2016.29.

[8] C. Yang, R. Harkreader, and G. Gu, “Empirical evaluation and new design for fighting evolving twitter spammers,” Information Forensics and Security, IEEE Transactions on, vol. 8, no. 8, pp. 1280–1293, 2013.

[9]Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, “Twitter spammer detection using data stream clustering,” Information Sciences, vol. 260, pp. 64–73, 2014

[10] Abreu, Jefferson & Ralha, Célia & Gondim, Joao. (2020). Twitter Bot Detection with Reduced Feature Set. 10.1109/ISI49825.2020.9280525.