Detecting fake news using geometric deep learning

Thomas Michel
thomas.michel1@ens-paris-saclay.fr
École Normale Supérieure Paris-Saclay
Gif-sur-Yvette, France

Bertille Temple
templebertille@gmail.com

ABSTRACT

As large social networks continue to thrive, the widespread dissemination of fake news presents a critical challenge, particularly during significant events like elections. Traditional fact-checking struggles to match the rapid pace and ambiguity of emerging information. This report examines the method proposed by [Monti et al. 2019], which employs Graph Convolutional Networks (GCN) for detecting fake news. We replicate and assess their model, investigating its performance on the FakeNewsNet dataset. Furthermore, we expand upon the feature ablation study outlined in the original paper, examining the significance of utilizing graph topology of the data in detecting fake news from social media.

Contents

1 INTRODUCTION/CONTEXT

In the era of the internet and social media, information travels at a breakneck pace, and concurrently, the speed of content creation continues to rise. These two phenomena render the fact-checking of such information nearly impossible, leading to the dissemination of misleading and false information (fake news). This issue has become more pronounced in recent years, especially during major political events, where fake news is used to sow confusion and manipulate the public. For example, a study by [Bovet and Makse 2019] suggests that approximately 25% of the tweets posted during the 2016 US election disseminated fake or highly biased news, potentially influencing public opinions about each candidate. The primary challenge with identifying fake news lies in the ambiguity of information, as the truth may not be widely available at the time of the news's emergence. Fact-checking can be time-consuming and necessitate rigorous journalistic or scientific investigations.

Various methods have been proposed for detecting fake news. At its core, the issue involves identifying a news story or a claim that can be proven false, meaning it can be verified and potentially disproven by thoroughly checking the actual facts. The initial requirement for a claim to be considered reliable is that it originates from a primary source, typically a news article that has supposedly gathered and verified the information it shares with the reader. The primary approach to detecting fake news involves attempting to assess the reliability of a news article or a social media post directly from their content.

However, the real threat of fake news arises when it spreads rapidly on social media, disseminating within a network of users who are often unaware of the validity of the claims and may struggle to evaluate the reliability of all the information they encounter

Previous studies indicate that genuine and false news follow distinct patterns of dissemination within a social network[Zhou and Zafarani 2020]. Specifically, false news tends to spread more effortlessly among users, eliciting stronger emotional reactions and higher reader engagement. Moreover, different types of news exhibit varying propagation behaviors across diverse communities. This gives rise to a second major research avenue for detecting fake news, which involves leveraging the circulation dynamics within the network and user information to differentiate between reliable and false information. The primary advantage of a propagation-based approach lies in its potential resilience to adversarial attacks, surpassing content-based methods. Simulating the writing style of genuine news appears simpler than replicating the entire dissemination pattern of authentic news. While the social context already furnishes valuable cues for identifying fake news[Nguyen et al. 2020], coupling it with the study of propagation patterns provides a more comprehensive understanding of how information spreads within the network.

Both previously described approaches rely solely on the information provided to users in a given news article or social media post but cannot ensure the veracity of the information. The last and most reliable method to detect fake news is to fact-check it by cross-referencing different sources or investigating the truth independently of the possibly unreliable information already available online. This last method is also the most complex, as it necessitates interaction between multiple modalities of verifying information and involves human experts, making it impractical for automated fake news detection.

In this report we will focus in details on the work of [Monti et al. 2019] which falls under the second category of methods and proposes a Graph Convolutional Networks for Fake News detection (GCNFN). We begin by explaining the approach presented by [Monti et al. 2019] and replicate their model. Subsequently, we replicate the findings of [Dou et al. 2021] concerning [Monti et al. 2019] model using the FakeNewsNet dataset. We then analyze the

impact of each provided piece of information on the model's accuracy. Going beyond the feature ablation studies conducted in both previous works, we assess a model that completely disregards features and solely considers the geometry of the data. Additionally, we evaluate a model that ignores the geometry and only utilizes the features associated with each node in the propagation cascade. Our demonstration reveals that the latter approach not only matches the accuracy of the model under investigation but can even surpass it with minimal modifications.

2 RELATED WORKS

Content-based fake news detection. News content-based approaches to estimate the reliability of news primarily attempt to extract features from the article or post that could characterize fake news. Certain writing styles and linguistic features can be identified as characteristic of potentially manipulative authors and baseless claims. Techniques such as the ones proposed in [Ghosh and Shah 2018; Kaliyar et al. 2021] employ natural language processing to automatically detect and extract relevant features to distinguish fake news from true news stories. However, this type of technique may face challenges in accurately detecting fake news beyond deceptive writing styles, prompting an exploration of detection techniques that emphasize propagation characteristics.

Network-based fake news detection. The goal of this line of work is to early identify fake news based on the way information spreads in a network. This idea stems from the observation that fake news propagation appears to be faster and more extensive in a social network. One hypothesis for this phenomenon is that fake stories evoke a more intense emotional reaction from the reader, leading them to share the information more readily. Another notion is that fake news has a greater impact in communities that are prone to sharing seemingly shocking information or news that reinforces their beliefs, facilitating deeper penetration of the fake information into the network. The use of various means of interactions provided by social media, such as comments and reposts, allows us to construct a graph describing how information spreads in a network of users; we refer to this graph as a propagation cascade. Assuming there is a difference in propagation between fake news and true news, we should be able to distinguish cascades resulting from fake news from those originating from true ones. Recent approaches[Dou et al. 2021; Han et al. 2020; Monti et al. 2019] employ graph neural networks to harness the geometric aspects of the spreading pattern, along with information pertaining to users and the content of their messages.

3 METHOD

The method proposed by [Monti et al. 2019] considers a spreading pattern as input, corresponding to the dissemination of a piece of news within a social network. A spreading pattern, or cascade, typically begins with an initial tweet containing a link to a news article. This tweet serves as the seed from which a tree-like structure of retweets and comments emerges. Fact-checking organizations such as PolitiFact[1] or GossipCop provide databases that link a set of articles to a brief statement and its corresponding label ('true' or 'false'). Some claims have ambiguous labels like "mixed," but these were filtered out in the studied article. Thanks to this database, the news referenced in the seed tweet can be associated with a brief statement and assigned the corresponding truthfulness label. The cascade generated by the seed tweet is then assigned this label.

Structure of the data. The model takes as input a news cascade structured as a graph. A graph consists of a set of nodes (tweets) connected by edges representing relationships between the tweets:

- Each node is represented by a feature vector. The model aims to integrate various approaches to detect fake news, resulting in node features that encompass heterogeneous data. This includes information about the social context of the tweet's author (such as the date of account creation), details about the tweet itself (like a word embedding of the textual content), and information about the spreading pattern (for instance, the retweet timestamp).
- A features vector of dimensions 4 is associated with each edge of the cascade. An edge is formed between node i and node j when the author of tweet i follows the author of tweet j, and vice versa, or if the author of tweet i retweeted the tweet from the author of tweet j, and vice versa.

Model. The input of the GCNFN model is a matrix that concatenates all the feature vectors of the nodes in the graph. For clarity in this explanation, the process of splitting the data into batches is not considered. The output of the model is a 2-dimensional vector representing the probability that the news is fake or true. Since binary classification is performed, the second probability may be deemed superfluous, but we chose to retain it for most experiments to stay closer to the original model. No performance cost was observed by either keeping or removing it. The implemented model consists of 2 graph convolutional layers and 2 fully connected layers, as depicted in Figure 1.

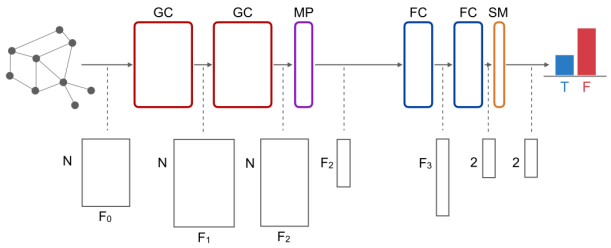

Figure 1: Model architecture. GC stands for Graph Convolution, MP for Mean Pooling, FC for Fully Connected, SM for SoftMax (Figure from [Monti et al. 2019])

Graph convolution. While information about the graph structure does not appear in the input tensor, it is included in the model through graph convolution. Convolution allows the integration of local neighborhood information into each node's representation, similar to how standard convolution in images incorporates local pixel information.

In contrast to the fixed number of neighbors for a pixel in an image, the number of neighbors for a node in a graph is variable,

[1] https://www.politifact.com/

and they lack a fixed relative position, making them indistinguishable from each other. This necessitates adapting the convolution operation to graphs.

A graph convolution entails aggregating features from node neighbors to update the node's own features. This is achieved by computing a weighted sum of the features of neighboring nodes, with the weights learned throughout the network.

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} c_{ij} h_j^{(l)} W^{(l)}\right).$$

A normalization constant is associated to each term of the sum:

$$c_{ij} = \frac{1}{\sqrt{|N(i)|}\sqrt{|N(j)|}},$$

with $|N(i)|$ being the number of neighbors of i.

The updated feature of each node becomes a function not only of its own features but also of those of its neighbors, reflecting the interconnected nature of the graph.

Attention. Instead of using a fixed normalization constant, the network can learn the relative importance of each neighboring node through an attention mechanism for graph convolutions introduced by [Veličković et al. 2017]. In the model under study, one attention head is included in every convolution layer.

Initially, the same linear transformation is applied to every node's feature to obtain Wh_i for each node i. Then, an attention coefficient (indicating the importance of node i for node j) is computed when there is an edge between these nodes, and we obtain

$$e_{ij} = a(Wh_i, Wh_j),$$

where a is typically a linear transformation of the concatenated vector. Finally, the attention coefficient is obtained by normalizing e_{ij} to make coefficients easily comparable across different nodes

$$\alpha_{ij} = softmax_j(e_{ij}).$$

Note that a non-linearity is introduced (Scaled Exponential Linear Unit or SELU in the model under study), so the attention weights are computed from $SELU(e_{ij})$ instead of dealing directly with e_{ij}.

At layer $l+1$, with $N(i)$ representing the neighboring nodes of i, the feature vector for node i can be expressed as:

$$h_i^{l+1} = \sum_{j \in N(i)} \alpha_{ij} W(l) h(l) j.$$

4 EXPERIMENTS

We conduct three main experiments:

- First, we investigate the influence of different features by employing a graph convolutional model and systematically removing certain features to observe the impact.
- Next, we compare a graph convolutional model that integrates various types of information (such as news spreading patterns, user details, and historical data) against a model that solely relies on news spreading patterns.
- Lastly, we seek to evaluate the significance of the graph's topology, leveraged through graph convolution, in identifying fake news. This is done by constructing a similar model that disregards the edges of the cascade and measuring its performance.

4.1 Experimental Setup

	[Monti et al. 2019]	[Dou et al. 2021]	Our experiment
Dataset	not published	politifact + gossipcop	gossipcop
Edges have features ?	yes	no	no
Train+validation -Test split (%)	80-20	20+10-70	20+10-70
Loss	HingeLoss	Binary cross entropy	Negative Log Likelihood
Activation	SELU	SELU	SELU
Optimizer	AdamGrad	Adam	Adam
Metric	ROC AUC	Accuracy and F1	Accuracy

Table 1: Comparison of experimental settings

Dataset. We assess the model using the FakeNewsNet dataset provided by [Dou et al. 2021]. This dataset comes in two variants, Politifact and Gossipcop, depending on the source used for fact-checking the news shared at the root of the cascades. For our experiments, we exclusively utilize the GossipCop dataset instead of Politifact, primarily due to its larger size. GossipCop comprises 5,464 graphs, in contrast to Politifact's 314 graphs, resulting in more meaningful results for our analysis.

We have access to two features: one containing information about the user profile (dimension 10), and another providing details about the content of the tweets posted by the author (dimension 300). These features can be encoded using the pretrained language model BERT or Word2Vec.

Models and evaluation. Since we only had access to data from Dou's research, our objective was to replicate their experiments. Consequently, we utilized the same loss function, namely cross-entropy. It's worth noting that in the context of a binary classification model with two classes, where the model outputs log probabilities for each class and a softmax layer is applied, the Negative Log Likelihood becomes equivalent to binary cross-entropy. Moreover, we adopted [Dou et al. 2021]'s unconventional data splitting approach and maintained the use of accuracy as our performance metric.

Accuracy represents the proportion of correct predictions made by the model out of all predictions. This metric is relevant in classification tasks, providing a straightforward measure of how often the model correctly classifies an input.

Feature\Model	GCNFN	MLP+MP	MLP+GAP
Content(W2V) + profile	0.97	0.95	**0.98**
Content (Word2Vec)	0.91	0.94	0.96
Content (BERT)	0.89	0.96	0.96
Profile	0.89	0.90	0.91
No feature	0.75*	/	/

Table 2: Accuracy results depending on the features used during training (MP=Mean Pooling, GAP=Global Attention Pooling, * Graph convolution without normalization)

4.2 Ablation study

In our study, we replicated the ablation analysis from [Dou et al. 2021], focusing on evaluating the impact of different features.

We were particularly struck by the high accuracy rate of 0.89 achieved using just a 10-dimensional user profile feature. This feature includes elements such as a user's verification status, their total number of tweets and likes, and whether they use geolocation, among others. However, it doesn't include any data about the tweet's actual content. The fact that we can classify a tweet cascade so accurately without considering the content of the tweets themselves is quite remarkable. That being said, it is crucial to highlight that although the graph structure is not directly present in the input feature vector, it is implicitly considered by the model during the graph convolution operation. In other words, the achieved accuracy of 0.89 takes into account not just the user profile feature, but also the graph's structure. In our second experiment, we plan to extend the ablation study beyond what was done in the article by removing the profile feature entirely and focusing solely on the graph structure information. Conversely, in our third experiment, we will remove the graph structure information, thereby isolating and evaluating the impact of the user profile feature alone.

Additionally, our findings showed that combining different types of information, like the user profile and the tweet content embedding, resulted in a significant increase in accuracy, jumping to 0.97. This is a substantial improvement over using either content or profile information alone, which yielded an accuracy of 0.90. This result aligns with [Monti et al. 2019]'s intuition about moving towards a unified framework that incorporates heterogeneous data.

In the end, our experiments comparing different encoding methods, such as Word2Vec and BERT, led us to conclude that the encoding method has a relatively small influence on the outcome.

4.3 Learning from geometry only

The original paper introduces a unified approach by simultaneously considering the content of each tweet, the social context of users, and the spreading pattern (the geometry of the cascade) to identify fake news. The authors emphasize that incorporating different types of information is crucial for improved results.

While previous approaches have examined the performance of content-based techniques, we aim to address the question of what we can learn from the geometry of the news cascade alone. Multiple approaches can be considered, such as building a classifier from handpicked statistics about the geometry of the spreading graph. Here, we chose to stay close to the original model and employ a similar architecture based on graph convolutions.

We constructed a feature-less dataset from the graphs contained in the FakeNewsNet dataset. To each node, we associate only one scalar feature, the same for every node. Due to the nature of the Graph Attention Layers, such data would always result in the same output. Therefore, to leverage local geometries of the graphs, we replaced these layers with simple graph convolutions as introduced earlier, without normalization coefficients (so the results of the operation also depend on the number of neighbors).

The results of this model, displayed in Table 2, show that we are able to obtain a significantly better-than-random classifier. This supports the original assumption of the difference in spreading patterns between fake news and true ones. As announced by [Monti et al. 2019], the unified approach allowed for improved performance compared to the geometry-only approach (at least for the proposed architecture).

4.4 Learning without geometry

While it may seem unconventional when dealing with data corresponding to the spread of news on a social media platform, one may wonder if it's possible to learn without utilizing the geometry of the spreading pattern. In other words, is the geometry of the news cascade relevant for fake news classification?

To address this question, we built a new model by replacing the graph convolution operations of the original model with linear layers applied to the features of each node independently, followed by a SELU activation. This implies that the features from neighboring nodes are not used when computing the new feature activation of the node. We derived two models from this idea: one where only the graph convolutions are modified, and another in which we also replace the average pooling layer with a global attention pooling layer introduced by [Li et al. 2015]. This pooling method allows for computing a weighted average of the features of the node, where the weights are determined by the attention mechanism.

This approach yields surprising results. As shown in Table 2, both non-geometric methods (MLP+MP and MLP+GAP) allow for the recovery of similar or better results compared to the approach proposed by [Monti et al. 2019]. In particular, we've identified that the best results can be achieved by using the combination of global attention pooling and a set of features comprising content and profile information, reaching an accuracy of nearly 98%.

This experiment demonstrates that the model based on geometric deep learning may either struggle to adequately leverage the geometry of its input or that the topology of the news cascade is not as crucial as the content of the tweets and the characteristics of the users involved. We believe that, with only two convolutional layers and average pooling, the model may not fully exploit the geometry of the cascade. However, due to the limited amount of training data at our disposal, it is challenging to provide a definitive answer. Indeed, adding more parameters to the models quickly leads to overfitting, perfectly fitting the training dataset but yielding poor results on the test set.

Although considering the spreading pattern in the classification task using graph neural networks results in similar performance as ignoring the graph structure. The real benefit of using a geometric deep learning model may not lie in improved performance, but rather in potential robustness to adversarial attacks and changes over time. However, the study of these properties is beyond the scope of our project.

5 DISCUSSION

5.1 Model and Reproducibility

The absence of published data from the original article [Monti et al. 2019] makes it challenging to replicate the study's findings. However, we successfully reproduced the experiments of [Dou et al. 2021] based on the published FakeNewsNet dataset. The quality of the results varies significantly with the quantity of data in the training set.

In particular, we found it surprising that the standard train-val-test split allocated the majority of the data to the test split (70%). The issue arising from this shortage of training data became most apparent when dealing with versions of the dataset with the highest number of features. This led to early overfitting of the dataset, halting improvements on the validation and training datasets. This probably also explains the relatively shallow architecture adopted by [Monti et al. 2019]. Indeed, adding more layers would not significantly enhance the model's performance due to severe overfitting.

The reproduced model, the variants as well as the code to reproduce the experiments are available on our GitHub repository[2].

5.2 Feature diversity against adversarial attacks

While the proposed model in the introduction of [Monti et al. 2019]'s paper adds value through its ability to handle diverse information, the paper does not explore robustness to adversarial attacks. In our project, we couldn't assess robustness against adversarial attacks due to the limited access to original social media data. Below, we provide some considerations on why the Graph Convolutional Neural Networks approach may prove more robust against adversarial attacks compared to our model, which disregards the geometry of the graph but achieves better accuracy.

It is known that a substantial number of social media accounts are owned by fake users and are used to automatically disseminate disinformation on a large scale. With recent developments in automatic text generation, the creation and dissemination of fake content become increasingly easier. Content-based fake-news detection approaches alone are likely to fall short against generative models in a prospective tug-of-war between fact-checkers and fake-news spreaders.

A propagation-based approach should be more resilient to attacks, as modifying the cascade's topology adequately for misclassification may prove challenging. Regular user interactions must be considered, and executing such an attack would require substantial infrastructure.

Even if the difficulty of fooling the model increases, making a fake news detector trigger at true news may not be as challenging, given that fake news propagation often already involves bot accounts. Thus, adding bot account activities in the news cascade of true news may confuse the detector.

Generalizing fake news detection models on social media platforms could be employed to attack true news, making them appear less reliable while promoting fake news. Therefore, an important aspect of fake news detection models is their robustness and resilience against adversarial attacks. It is crucial to consider multiple types of information at our disposal—content and propagation cascade, and perhaps the reliability of sources—to build a resilient model. This is why a model that ignore the geometry of the propagation graph may not be wise with regard to robustness.

6 CONCLUSION

In this project, we successfully replicated the findings of [Monti et al. 2019] using a different dataset proposed by [Dou et al. 2021]. The results align with both papers, showing that the studied method achieves effective fake news detection using both the content of social media posts (specifically Twitter) and the cascade structure originating from the original news post. We extended the ablation study conducted in the original paper by exploring different encodings for the dataset and considering cases where we disregard the cascade's topology.

An intriguing result emerged from the experiment where we excluded the cascade's topology. Aggregating the results of the content of each post in the cascade through an MLP yielded performance close to the original architecture. This suggests that the exact cascade topology may not be as crucial as the content of user responses, provided a sufficient number of related posts are considered. While highlighting the importance of robustness against attacks for real-world deployment, we acknowledge a lack of understanding regarding how different approaches may be affected by adversarial attacks.

Through this study, we identified several avenues for further exploration. First and foremost, we emphasize the need for a comprehensive examination of robustness against adversarial attacks and when faced to new data overtime (concept drift). The technique we replicated has seen multiple studies[Han et al. 2020; Horne et al. 2019] in this direction, and we suggest that this line of work should be extended to develop more robust algorithms. Additionally, alternative approaches, such as the one without graph convolution mentioned in this study, should be considered to understand the impact of this simpler architecture on the robustness of the model.

Moreover, we would like to mention some minor improvements upon the studied technique for future works. Firstly, optimizing the content encoding specifically for fake news detection could enhance results. While our study used a dataset with premade encodings, fine-tuning the encodings for the task could likely improve performance. Another potential improvement is to treat fake-news detection as a node classification task instead of graph classification. This approach would offer more flexibility in classifying social media messages. In the considered propagation graph, not all users necessarily support the fake news, and some may aim to reestablish the truth. Therefore, we may want to develop a more fine-grained approach by adjusting the current method.

Finally, although the evaluated methods in this report perform very well on their task with the given dataset, they lack real-world context. Additionally, we classify news as either true or false, while reality is often more nuanced. Perhaps a tweet is just misleading, maybe there isn't enough information to conclude, or maybe the model classifies in a certain way, but the clues leading to the decision are mixed. We believe that a reliable system to detect fake news should be able to justify its decision, provide an estimation of its confidence in the result, and sometimes withhold a hasty judgment that could be detrimental to the involved user.

[2]https://github.com/Thomick/Geometric-fakenews-detection

REFERENCES

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 7.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2051–2055.

Souvick Ghosh and Chirag Shah. 2018. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology* 55, 1 (2018), 805–807.

Yi Han, Shanika Karunasekera, and Christopher Leckie. 2020. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint*

arXiv:2007.03316 (2020).

Benjamin D Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 1 (2019), 1–23.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications* 80, 8 (2021), 11765–11788.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* (2015).

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* (2019).

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1165–1174.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.