Reviewing Fake News Detection on Social Media using Geometric Deep Learning

Prakhar Agnihotri
2017A8PS0280P
BITS Pilani
Pilani, India
f20170280@pilani.bits-pilani.ac.in

Pratik Borikar
2017B3A70550P
BITS Pilani
Pilani, India
f20170550@pilani.bits-pilani.ac.in

Niranjan Ashok Jahagirdar
2017B3A70454P
BITS Pilani
Pilani, India
f20170454@pilani.bits-pilani.ac.in

Aditya Saxena
2017A7PS0166P
BITS Pilani
Pilani, India
f20170166@pilani.bits-pilani.ac.in

Manjot Singh
2017A1PS0662P
BITS Pilani
Pilani, India
f20170662@pilani.bits-pilani.ac.in

***Abstract*—This is a report to review the paper, "Fake News Detection on Social Media using Geometric Deep Learning". Through this report we will describe the problem statement, past research methodologies, major contributions, algorithm architecture, limitations and future research directions with respect to the aforementioned paper.**

***Index Terms*—review, fake news, geometric deep learning**

I. PROBLEM STATEMENT

The primary problem the paper addresses is the one of Fake News Detection on Social Media using methods of Geometric Deep Learning. Authors aim to find robust mechanism of detecting fake news as a means of countering propaganda and disinformation campaigns. The mechanisms to be developed should be independent of features like language, locales, geography etc. as to ensure working in a general sense i.e. independent of these features. It is essential to ensure that the developed method is robust enough to remain unaffected by manipulation from external sources and adversarial attacks. Finally, it is to be ensured that the method ages well over time, and works even after long periods of time after training.

II. IMPORTANCE OF THE PROBLEM

There is sufficient evidence that with the increasingly open and free access to the internet, use of social media has spread like wildfire. More importantly, due to rapid circulation of news on these platforms, social media has become the sole source of news for most people. With the increase of news sharing channels on social media, it becomes increasingly important to identify and verify the information that is shared on these platforms. Beginning with the Cambridge Analytica Scandal during the 2016 Presidential Elections to manipulate public opinion, the importance of social media in shaping people's opinions has come to the forefront. Similar manipulation was seen during the Brexit referendum, and even now, during the COVID-19 pandemic, with ongoing Russian disinformation campaigns. Hence, finding a solution to this problem has become increasingly important for the benefit of society as a whole. It is to be noted that task of verifying news is difficult for general public since most of the fake news on social media requires social or political context. Hence, there is a need of a mechanism that can efficiently distinguish fake news from the truth.

III. PAST RESEARCH DIRECTIONS

Past research could be divided into three categories based on which approach they use to identify fake news.

A. Content Based

These approaches rely on linguistic characteristics, to capture indications that may be associated with fake news. Content-based approaches are content, and hence language dependent which restricts their use in generic approaches. A major drawback of this approach is that fake news can be labelled as true news if it's well written and does not appear as fake at first sight.

B. Social Context Based

These approaches use context features such as demographic information, social network structure including connections, friends, followers etc. and reactions by users (likes) to identify and generalise fake news sources. The social context during news dissemination process on social media forms the inherent tri-relationship, the relationship among publishers, news pieces, and users, which has the potential to improve fake news detection [1]. However, this is found to not work very well on its own.

C. Propagation Based

These approaches track the spread of news over time to find features distinguishing the propagation pattern of true news versus the pattern of fake news. It is based on the assumption that news spreads in a manner similar to that of infectious

diseases and hence, we can use epidemic models here too. It has been empirically found that the spread of fake news is different from true news. This allows for a model to be built with this in mind, promoting a very general approach, agnostic to the content features like language, geography etc. These approaches are possibly robust to manipulation from actors spreading misinformation, since spread is not controllable by these actors, therefore propagation pattern cannot be manipulated. However, past approaches have focused too much on graph theoretical features while designing models which seem to be too general for the specific task of FND.

IV. MAJOR CONTRIBUTIONS OF THIS PAPER

Exploited geometric deep learning techniques to identify features that could possibly help distinguish between true and fake news. Moreover, most of the research in the past with regards to Deep Learning approaches such as CNNs, LSTMs etc have made the tacit assumption that the data is grid-structured or Euclidean. Hence, this is one of the few approaches to use non-Euclidean data such as graphs, in this case, successfully to actually solve a real world problem. What makes this approach different is that any attempts until now, to use propagation features had focused on theoretical aspects of graphs, such as cliques, degrees, edges etc., which was not found to work well since they were not very task specific, and very general. In this paper, GCNNs were used so that propagation data could also be incorporated. These work on graph structured data, and can incorporate heterogeneous data (such as user demographics, activity, social network structure, propagation structure and even content), hence allowing for a very holistic approach to identifying fake news. What is more groundbreaking, is the fact that the approach used by the others is neither solely context-based nor content based, but uses propagation of the information as the most important factor in identifying the validity of the news. Such an approach, is conjectured to be extremely robust to any kind of adversarial attacks by actors who would wish to cheat the system, since news spread is not easily controllable by them.

Paper describes a URL or a cascade arising from a URL with corresponding tweets by using graphs with tweets in each, (acting as nodes) with estimated news diffusion paths and social relations represented as edges. The importance is how two models were proposed on seemingly similar, yet distinct features. The cascade is the news diffusion tree that comes into existence due to a source tweet that references an URL, along with all its retweets. With just these two variations important differences in model performance were identified. In such an approach, GCNNs make use of the Eigen-decomposition of the graph Laplacian Matrix to process non-Euclidean data. The Adjacency matrix is taken into account along with the relevant node features we pass to the function. With features present in the Feature Matrix and node connections (social structure, in our case) present in the Adjacency Matrix, this allows the proposed model to learn the features and spread that differentiates the two outputs.

This paper uses fake and true stories (verified by Snopes and Politifact) spread on Twitter between 2013-2018, for training, eventually achieving a 93% ROC-AUC, even with very short spread times ($\tilde{2}$ hours of propagation). Moreover, the model performs extremly well even when the model is trained on data that is distant in time for the test data.

It was observed that credible and non-credible users tend to form two distinct communities, and the ”tweeters” seem to have mostly homophilic interactions. This is quite similar to the socio-political concept of echo-chambers. Moreover, the paper seems to also score each user, on the basis of their credibility, between -1 and +1, and computes it as the difference between the proportion of retweeted true and fake news. Why this is important, is that it may allow future researchers to use these scores in other applications, and might also pave a way to actually improve people's critical thinking when it comes to them identifying fake news. Hence, the ramifications are immense, which could possibly improve the entire society as a whole, decreasing the spread of fake news, allowing people to protect themselves from nefarious actors who may otherwise take advantage of their gullibility.

They also found that all features positively added to final predictions, except tweet content, removing which improved performance by 4%. It was hypothesised that since 20% of cascades were related to 1.5% of the URLs, this might have led to over-fitting with regards to content.Hence, removing this feature, led to an increase in performance on the test set. This fact also goes to show how fake news spreads from only a few URLs and is able to spread to so many users just due to social media algorithms, which in itself is a boon and a bane.

Since a significant improvement in performance is visible on increasing the news spread time for data from 0 hr, it indicates and supports their hypothesis that the propagation based features play a significant role in distinguishing the two types of news, and there is a difference in their spread.

It was observed that the performance of the model drops after the age difference between the real-time data and the data on which the model was trained, exceeds 180 days for the URL-wise setting, and 260 days for the cascade-wise setting. Such a difference might be because cascades have higher variability, hence forcing the model to learn simpler and even less discriminative features, which leads to lower overall performance but more resilience to model aging.

V. LIMITATIONS OF THIS PAPER

The major limitations of this paper stem from the assumptions made by its authors and the data on which the network is trained and tested.

As has been mentioned, the authors rely on Snopes, PolitiFact, and Buzzfeed for the classification of true and false statements. The authors of [2] collected a random sample of 858 fact-checks and evaluated them in the light of criteria based on or inspired by fact-checking literature and the International Fact-checking Network's code of principles. Their analysis revealed that while PolitiFact fared well in general there is much room for improvement. This is particularly true

in case of complex propositions. where there more than one proposition. In 279 cases (33% of the sample), PolitiFact checks a complex proposition and assigns one truth rating to it. This isn't entirely accurate as some of the claims in a piece labelled false may actually be true. PolitiFact also checks claims that authors of [2] consider uncheckable. These are statements whose truthfulness cannot be defined in practice, e.g. claims about the future and vague claims. In 92 cases (11% of our sample), PolitiFact checked a claim like this. Therefore, when a model is trained on this data, there is a bias/error that is carried forward from the PolitiFact dataset.

This paper uses the findings of [3] to substantiate the claims of categorization of tweets but this has been confirmed only for political tweets. For the rest, there is no such empirical confirmation. As shown in [4], political fake news is only one type of fake news that is spread on fake news. In addition to this, we have fake news regarding scientific research, pranks and even active disinformation campaigns by actors to gain fame. There is no evidence of the categorization of such tweets the way there is for political tweets. Therefore, the model proposed in this paper makes an assumption of categorization that perhaps may not apply to non-political fake news which is as important and can be as fatal as political fake news. Hence, without enough evidence we cannot say whether this model will work well to accurately check the truthfulness of non-political tweets.

This paper uses a cascade based approach where the size of the cascades are defined as the number of tweets in a particular cascade. The average cascade size in the dataset being used is 2.79. Only cascades with 6 or more tweets were used for the cascade-wise classification. This was done because cascades of size less than 6 tweets did not display a clear diffusion pattern which is crucial to the classification. However, this led to the exclusion of 5,976 thus severely reducing the size of the dataset. Furthermore, the paper assumes that all cascades associated with a URL, inherit the URL label. This could be false when the particular tweet/post is denying content of the URL. The assumption that all the cascades associated with a URL inherit the label of the latter is a fallible assumption.

The model mentioned in this paper is too simplistic because it ignores mixed or partial true/false data, losing an important chunk of the set. Complex statements, as mentioned before, need not be completely true or completely false. In today's world there is often a spectrum when it comes to the truthfulness of statements which is used by some nefarious to intentionally mislead people. Furthermore, there are questions concerning the validity of the current trained model in future predictions given the change in nature of social media. The model seems to not age well with time, with performance drops drastically after 180 days in URL wise settings, and 260 days in the case of cascade wise setting.

VI. Future research directions

In this sections we will discuss possible improvements to address the limitations mentioned previously and we will also discuss applications of this paper which can be explored further.

As has been mentioned, while preparing the data for this paper, the authors have tried to inlcude only those statements that are completely true or completely false. However, often complex statements cannot be assigned a singular true or false label. Researchers can aim at increasing the number of categories by including partially/mostly true or false that shall let people know that this is a complex statement and hence, the model cannot label them definitively. Furthermore, since the model loses a lot of data by excluding small cascades, the model can make a hybrid of a content-based system for smaller and and propagation-based system, with robust regularization.

We could possibly combine the URL and cascade wise approaches, since the former has better overall performance but the latter is more robust to aging. Regularization is important, to get the model to perform well even on different test sets in terms of language, content and time in addition to experimental validation of the conjecture that the model is robust to geography and language.

We need to experimentally verify how robust the model is to adversarial attacks.The study of adversarial attacks is also of great interest, both from theoretical and practical viewpoints. The adversarial attacks would allow exploration of the limitations of the model and its resilience to attacks. It has been conjectured that attacks on graph-based approaches require social network manipulations that are difficult to implement in practice, making this method particularly appealing. On the other hand, adversarial techniques could shed light on the way the graph neural network makes decisions, contributing to better interpretability of the model. Researchers can also explore additional applications of our model in social network data analysis going beyond fake news detection, such as news topic classification and virality prediction. We can improve interpretability of model by understanding what features contribute to the difference between true and fake news, since proving it is a must, if it is to be adopted.

The model has been trained and reestricted to Twitter. However, researchers can look at replicating this model for other social media platforms like Facebook. This can be done by replacing the retweet metrics with the shares, no of reacts and the number of friends associated with a post.

VII. Implementation

The GCNN architecture implemented in our code follows the architecture given in the FND paper:

- The first GC1 layer implemented in our code is the GATConv layer, which is the graph attention operator, which is majorly implemented in graph classification problems. The size of each input channel is 4 (no of features) and the size of each output sample is 256 (2* hidden layer, where the hidden layer has size 128), where the SELU activation function is applied on the output.
- The SELU (Scaled Exponential Linear Unit) activation function is defined as follows for input X (element-wise):
  - X greater than 0, implies return lambda * x

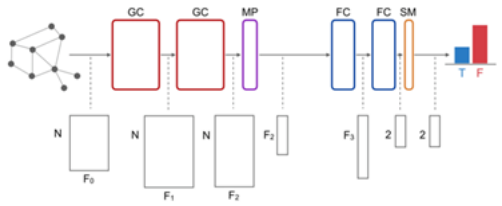

Fig. 1. Model architecture.

– Else, return lambda * alpha * (exp(x) -1), where scale = 1.05070098 and alpha = 1.67326324 by default

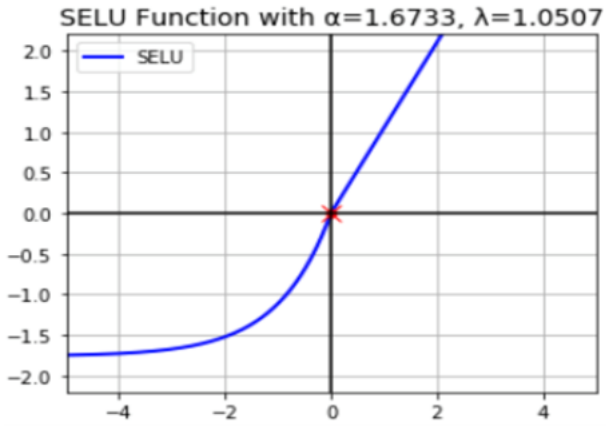

Fig. 2. SELU function

- The second GC2 layer is again a GATConv layer, but has an input size of 256 (2* hidden layer, where the hidden layer has size 128) and an output size of 256 (2* hidden layer, where the hidden layer has size 128) followed by the application of the SELU activation function giving us the output of GC2.
- The global mean pooling layer is applied to the output of GC2, followed by the SELU activation function.
- Now, the FC1 layer is a simple linear perceptron layer having the input size as 256 (2* hidden) and the output size of 128 (hidden layer) followed by the SELU activation function.
- The output of FC1 layer servers as the input for FC2 layer, which is a linear layer having input size of 128(hidden layer) and an output size of 2 (no. of classes, here it is 2 since it is a classification model).
- SoftMax activation function is applied to the output of FC2, giving us the probability of a particular input belonging to anyone of the two classes (0 or 1). The probability of the class which is higher is the final output of the neural network.For Example, if the probabilities are [0.5612, 0.4388] (Note: Sum will always be 1), the input would be classified as the one belonging to class 0, since it has a higher probability.

We have used the PROTEIN dataset for training and testing our model. In this dataset, a protein is classified as an enzyme (class 0) or a non-enzyme (class 1), where nodes of a graph are amino acids and an edge between them represents that the distance between two nodes is less than 6 Angstroms. The following shows the dataset details and other parameters used in the model:

- Features (number of types of nodes): 4
- Classes: 2 (0 for enzyme, 1 for non-enzyme)
- Average nodes per graph/data-point: 39.06
- Average edge per graph/data-point: 72.82
- Epochs: 100
- Hidden layer size: 128
- Loss function:Negative Log likelihood loss (NLL Loss)
- Optimizer: ADAM optimizer

The following graphs represents the results of training and testing on the dataset. Please note that the dataset has been split into training and testing samples in the ration 8:2.

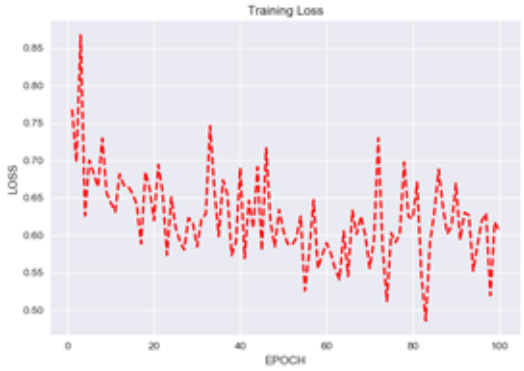

Fig. 3. Training Loss

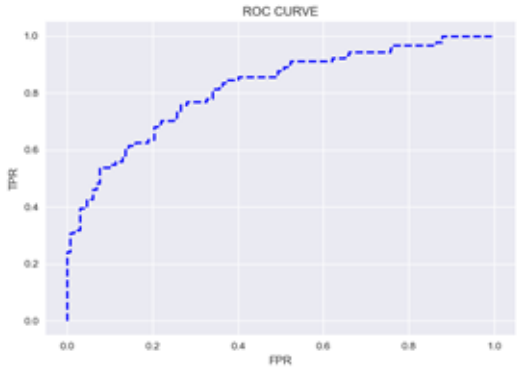

Fig. 4. ROC curve on test dataset

ACKNOWLEDGMENTS

We would like to thank Dr. Vinti Agarwal for giving us the opportunity to work on this project. We would also like to thank the teaching assistants, Aditya Deshmukh and Raksha

Chaudhary for being very patient with us and answering all our doubts.

REFERENCES

[1] F. Monti, F. Frasca, D. Eynard, D. Mannion and M. M. Bronstein. Fake News Detection on Social Media using Geometric Deep Learning. arXiv preprint arXiv:1902.06673, 2019.

[2] S. Nieminen and V. Sankari. Checking PolitiFact's Fact-Checks. Journalism Studies, 22:3, 358-378, 2021.

[3] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In Proc. ICWSM, 2011.

[4] Kalsnes B. Fake news. Oxford Research Encyclopedias: Communication published online September 2018.