Fake News: Feature Analysis and Detection - Interim Report

Anchit Bhattacharya, Ganesh Rakate
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
{abhatt22, grakate}@asu.edu

November 26, 2018

1 Introduction

Social media has made consumption of information low cost and very easy to access. However, with the ease of consumption of news, it has also brought about wide propagation of fake news i.e news which intentionally spreads false information (Shu et al. [2017a]). Fake News has negative impacts on an individual as well as on society. Therefore it has become increasingly important to build fake news detection models, but it's a technically challenging problem. Challenges in the Fake News Detection problem ranges from collecting reliable data with ground truth, extracting important features, and using these features to build effective models. Most of the data used by researchers for fake news detection has been based on the News Content, but recently new research has emerged where people are trying to use social context features which acts as an auxiliary information for fake news determination. One such data set is the FakeNewsNet Data which contains two data sets with news content and social context features (Shu et al. [2018b]). The FakeNewsNet data consists of news articles separated into fake and real news with each news article having information about the News Content, tweets related to the news, retweets, replies and likes related to each tweet. Also, it has information related to user profiles, their relations with the type of news (fake or real), and the follower and followee information for each user. This is a very rich set of data, which can be used for analyzing and extracting new features, visualize them and use them for creating novel models for detecting fake news.

The features for fake news detection can be classified into news content features and social context features. The news content features can be divided into Linguistic-based and Visual-based. There are three types of social context features, namely, user-based, post-based and network-based (Shu et al. [2017b], Shu et al. [2018a]). User-based features are characteristics of the users having interactions with news on social media. Post-based features are unique features of posts based on the social response of users, such as stance, topic and credibility. Network-based features involve constructing specific networks between users involved in the social media posts. There are different kinds of networks that can be constructed such as stance networks, co-occurrence networks, friend-

ship networks, credibility networks and diffusion networks. Each network has different type of node and edge information, and analyzing these networks on the Fake News Dataset can help bring out exciting insights about Fake News, which can help us in making better fake news detection models. Additionally, studying networks can help us in understanding Fake News Diffusion and Fake News Intervention.

We are interested in constructing various networks using the FakeNewsData set we have, and visualize/analyse these networks to find some insights. We would like to analyse as many networks as we can, and create visualizations, but due to the time constraint of the project and the complexity of the data, we might be limited on this aspect. Also based on the network analyses results, and their discrepancies on fake news and true news, we further aspire to create a model based on our findings of the network measures using these network features.

2 Problem Statement

As described in the introduction, our primary task is to take the FakeNewsNet dataset, and construct and visualize various networks, with the intention of using this network features for fake news detection model construction. In this section we provide more details on the structure of the dataset, as well as the mathematical models of the different networks we propose to model, to provide a more formal understanding of the problem we are trying to solve.

2.1 Structure of the dataset

The dataset (Shu et al. [2018b]) contains data from two fact checking websites, politifact and gossipcop. For our problem, we will focus on the politifact dataset. The data is divided into two folders, politifact real and politifact fake, which corresponds to real news and fake news. Each of these folders have subfolders relating to one news article, which contain the following five files:-

a. ***newsarticle.json*** - This is a json format file which includes text content of the news, image URLs, news publication date and publisher information.

b. ***tweets.json*** - contains tweet information related to the news

c. ***retweets.json*** - contains array of retweets for each tweet.

d. ***replies.json*** - contains array of replies for each tweet. Some reply object have multiple level of replies

e. ***likes.json*** - contains array of user ids who liked each tweet

Additionally, we have a userprofilemetadata.json for both fake and real news, which has information about user profiles of users related to each type of news.

Finally, we have userfriendsid.txt file, which contains the user id of each user, and the follower and followee information(represented by the twitter user id) for each of the users.

2.2 Network Types

These are various kinds of networks (Shu et al. [2019]) we plan to model:-

2.2.1 Friendship Networks

Friendship network for a user can be modeled as a directed graph, $G_F = (U, E_F)$ where U and E_F are node and edge sets, respectively. A node $u \in U$ denotes a user, and $(u_1, u_2) \in E$ denotes whether there is a social relationship between the two users. This type of networks helps in understanding social relationships between users, which can help in analyzing spread of news, and community information.

2.2.2 Diffusion Networks

A Diffusion Network for a user can be modeled as a directed graph, $G_D = (U, E_D, p, t)$ where U and E_D are the node and edge sets respectively. A node $u \in U$ can publish, receive, and propagate information at time $t_i \in t$. The directed edge, $(u_1 \rightarrow u_2) \in U$ denotes the flow of information from node u_1 to node u_2. The probability of information propagation from node u_1 to node u_2 is denoted as $p(u_1 \rightarrow u_2) \in [0, 1]$.

2.2.3 Interaction Network

An Interaction Network is a heterogeneous type of network. It can be modeled as $G_I = (\{P, U, V\}, E_I)$ and it's edges (E_I) show the interactions between the nodes news publisher (P), users (U), and the news articles (V). The interaction network describes the interactions between different entities. These interactions are important features to differentiate between real news and fake news.

2.2.4 Stance Networks

It's a type of heterogeneous network where the node entities can be of different types. It can be represented as $G_S = (U, S, V, E_S)$ where the nodes can be posts, news items and users and the edges E_S represents the link between two types of nodes like stance between a news item and post. Stances signify the viewpoint of an user towards a news, such as supporting, opposing, neutral etc.

Based on the above information, our problem statement is to create algorithms to process the data, feed it into network generating tools such as networkX to form the different networks mentioned above, and create relevant plots, visualize the data, and generate network measures, for each of these networks, and do a comparison analysis of each of these networks for fake news and real news dataset.

3 Proposed Method

3.1 Friendship Network

1. Import the friendship network data in Python, in a dictionary form, with key being the user id, and the value being a list of id's which represents the follower/followee.

2. Loop through the dictionary, to create unique node for each user id, and form a directed edge between each user and his follower, pointing from the followers to the users.

3. Similarly, create directed edge from a user to its followees, pointing from a user towards the followee.

3.2 Diffusion Network

1. For each tweet in the tweets.json, create a node for each user id related to the tweet. Create a directed edge, using the time-stamp of the tweet for each user, pointing form the user who tweeted before, to the user who tweeted later, only if the user who tweeted later is present in the follower list of the first user.

2. To check if a user is in the follower list of another user, make use of the friendship network data.

3.3 Stance Networks and Interaction Networks

We are still brainstorming algorithms for other networks such as stance networks, interaction networks etc. Once we are ready with the method we will update this section.

4 List of Tasks

1. Study Network Analysis and Data Mining Techniques for Fake News Representation (Anchit)

2. Study the FakeNewsNet data set structure. (Anchit and Ganesh)

3. Process the data into correct format for use in algorithms (Ganesh)

4. Create Algorithms to feed data into Network tools (Anchit)

5. Implement Algorithm in Python (Anchit and Ganesh)

6. Create different networks. (Ganesh)

7. Network visualization (Anchit)

8. Calculate the network measures. Study and compare network measures for different networks created. (Ganesh)

9. Analyse network measure results for Fake News and Real News Data Set (Anchit and Ganesh)

10. Writing the final report (Anchit and Ganesh)

Until now, we have completed the first two tasks and currently working on next two tasks.

References

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017a.

Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017b.

Kai Shu, Deepak Mahudeswaran, and Huan Liu. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, pages 1–12, 2018a.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018b.

Kai Shu, H Russell Bernard, and Huan Liu. Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pages 43–65. Springer, 2019.