users and repositories, along with the *followee-follower* relationship among users, forks and watch events between users and repositories to build our network structure. We developed a deep feedforward neural network model: "*DeepFork*"; a supervised machine learning based approach. Our DeepFork model is a direct contribution to SocialSim project and the code is available online[1]. DeepFork is trained with both node and topological features of network to predict information diffusion. To evaluate the performance of DeepFork, we compare it against the benchmark machine learning classifiers and DeepFork performed fairly well. Moreover, we perform an ablation study by ignoring the watch events, to inspect if user awareness in "Watch"-ing the repository helps in predicting whether or not that user will fork that repository.

This paper is organized as follows: related work is detailed in Section 2. Section 3 explains, the proposed approach along with problem formulation, datasets and data preparation, feature extraction and the model description. Section 4 presents, the experimental setup for empirical study using performance metrics. Section 5 details the results and analysis of the developed model, followed by the conclusion and future scope for extension in section 6.

2. Related Work

In this section, we discuss the related work in areas of link prediction, information diffusion and supervised machine learning and also analysis on Fork events in GitHub.

2.1 GitHub

Although GitHub is originally launched for software development, through millions of repositories that are freely available, it has also become a knowledge sharing resource around the world for both industry and academia. Forking is usually the first step towards contributing to a project, by modifying a copy of the original project. By the fork event, developers will have a privilege to access the code and use it as and when needed. Based on the study conducted by (Jiang et al. 2017), developers perform the fork event mainly to contribute to the original repository primarily based on their programming language preference. Based this study, to address why and how developers fork, (Zhang et al. 2017) showed that the recommendation of repositories based on user preference has made a significant difference in the overall software development trends, while the open source software profited the most. These recent studies motivated us to learn more about how a piece of information is diffused among assorted repositories and users in GitHub.

2.2 Information Diffusion Prediction

We adopt definition of diffusion of information from (Rogers et al. 2010), i.e., the transfer of information from one entity to another that can happen through certain channels over a period of time, irrespective of that information being novel or not. There are many existing models in literature for information diffusion, but for this work, we focus on machine learning based link prediction models as well the models that use network structure for link prediction.

2.2.1 Network Structure-based Link Prediction:

From past few decades, another emerging line of study in complex social networks is link prediction, ranging from node pair similarities to learning the features of the nodes itself. This kind of study is grounded on a simple logic that any two nodes with compelling similar properties tend to have an interaction between them. When we think of user similarity and social networks, the first thought that occurs to the mind is the user profile. (Bhattacharyya et al. 2011) and (Anderson et al. 2012), conducted an experimental study to acquire knowledge on user similarity by analyzing user profiles and their corresponding interests groups. The actions performed such as editing the article, answering the questions by those users in Stack Overflow and Wikipedia, are recorded into a vector and then cosine similarity between the vectors are calculated to get the node pair match. In extension to this, (Akcora et al. 2013) regulated an exploratory analysis and believed that the actions performed by the users and their corresponding attributes can reflect the personal interest and social behaviors, and hence there is a likelihood of new connections to form among those users. On the other side, the study on topology also gained focus after (Nowell and Kleinberg 2007) graph structure-based work by analyzing the "proximity" of nodes in a network for link prediction. Metrics used for link prediction that is purely based on the topological information are classified into neighbor based, path based and random walk

[1] https://github.com/akula01/DeepFork