Joyce Bozeko
joyceb@vt.edu
Social Media Analytics Final Project

Abstract:
Reddit Hyperlink Network is composed of complex connections of webs between subreddits which are created through the hyperlinks contained in posts. Relationship between subreddit is represented by the directed network, post is the source subreddit which contains hyperlink and subreddit is the target subreddit which points the hyperlink. Aim of this research is to perform the link prediction on the dataset of reddit hyperlinks and to analyze the dataset to find out the potential links which exists between two nodes. To perform this analysis useful features will be extracted from the dataset, those relevant features could be used to predict the likelihood of a link existing between two nodes. Dataset of reddit hyperlink is ranging from January 2014 to April 2017 and total number of nodes in the dataset are 55,583 nodes and the edges between these nodes are 858,490 this is enough data to perform the link prediction on hyperlinks. To gain insights about the structure of reddit hyperlink network and to identify which key subreddits perform a major role in connecting with different communities on the platform this data has enough information to carry out this case study. Link prediction is based on identifying the common neighbors between two nodes, their similarity of node attributes and the distance of these neighbors in the network. Feature of common neighbor's measures how many subreddits that are already to the two nodes, to access the similarity of nodes the attributes of two nodes are compared. Shortest path between two nodes in the network is measured by the distance feature. For evaluation of the link prediction made by model different range of metrics will used including precision, recall and F1 score to validate the prediction made by model. To validate the model accuracy random data will be passed to model to predict effectiveness of model.

Introduction:
Advancement in technology and excessive use of internet has given rise to the development of social networking sites to connect with people of same interests. Reddit is a famous social media platform where subreddits are created by users to discuss about specific topic. Subreddit is an online community of users of same thoughts where they discuss about their topics including sports, politics, technology or any other topic they want to discuss. The Reddit Hyperlink Network dataset shows the data of the links created in posts of users from various communities on the website. Each hyperlink in the dataset represents the source to its target. The subreddit from which the linked-to post originated is known as the "source subreddit," whereas the "target subreddit" is where the link actually leads. The dataset spans approximately 2.5 years, from January 2014 to April 2017, and includes around 55,863 nodes (subreddits) and approximately 858,490 edges (hyperlinks between subreddits). Network analysis relies heavily on the prediction of link which helps to understand the link between two entities. This link prediction can help can be useful in many other tasks such like recommendation systems if you are able identify that what kind of content or product is used by two people you can recommend the same product to other people who have same attributes, social network analysis can help to identify the social influence. Social networking site like twitter where users can follow each other, and they retweet the tweets of others our goal is to identify those users who make significant impact on the behavior and the attitudes of other. Link prediction can help to identify those users based on their retweets, mentions and other relevant features. Link prediction can be used in marketing to find out the potential customer. By analyzing the link between the customers and products. By using link prediction, the target can be easily analyzed, and the marketing strategies and customer segmentation can be adjusted according to that. Link prediction finds out the likelihood of a link between two nodes based on the observed structure of the network. To carry out the link prediction the useful features are extracted from dataset which can be used to find out the likelihood of two nodes with each other. Common neighbors between two nodes (number of subreddits that are already associated to the two nodes), distance of nodes in the network and similarity in the attributes of the nodes are used as features to perform link prediction.

In networking analysis, link prediction is the main topic which has been carried extensively and many solutions has been proposed on this topic. Measuring the degree of similarity between pairs of nodes in a network is a popular approach. The nodes in the network that are similar to each other are considered as more likely to be connected in the network. Another approach is used in link prediction to find out the adjacent nodes. The node which highest number of adjacent nodes is considered as to be connected in the network. The main behind this approach is that a node which highest number of adjacent nodes are likely to have more opportunities to form a link in the future [1]. The aim of this research is to apply link prediction on the dataset of the reddit hyperlink network. The data of reddit hyperlink will be passed to machine learning algorithm to perform the training for the prediction of the link. Once the machine learning model get trained on that the model will be evaluated using different evaluation metrics to validate the effectiveness of the prediction of link. Last but not least the reddit hyperlink network dataset allows to perform the link prediction to find out the understanding of reddit platform how people are connected with each other. This project can enable many other findings which can be used for marketing, customer segmentation, recommendation systems and social network analysis.

Background:

In this section background information of the link prediction is going to be discussed. Current attraction of the social network will be discussed first then till what extent link prediction can be used in social networks will be discussed.

- **Approaches of Link Prediction in Social Networks.** Link prediction is the major task which is carried out recursively in social network analysis. Different approaches for link predictions are used each approach has its pros and cons. One of the famous approaches of link prediction is prediction of link based on the similarity. This approach considers the nodes that are similar to each other they are more likely to be linked together. Different similarities measures are used to find out the similar nodes including Jaccard index, cosine similarity and Pearson correlation coefficient. This approach finds out that which node has highest number of adjacent nodes, and that node are most likely to be linked. Second approach of link prediction is based on the path. This approach assumes if any node is connected to the shortest path or high-quality paths are more likely to be linked than nodes which are connected to longest and low-quality paths. Different measures are used to compute the length and quality of the path including such as Katz index, Adamic-Adar index and preferential attachment. These measures compute the quality of the nodes if they node has the highest quality are more likely to be connected path. Third approach of link prediction is based on the community. This approach finds out if any node that belong to the same community or having same structure of node from the community that node is more likely to be linked. If a node has structure which don't resemble to the node from a node of any community that doesn't consider to be the connection. Different measures are used to compute such as common neighbors from community, probability of nodes that belongs to same community and community-based similarity measures. These measures compute and find out if node has the same attributes from the node of community if the find out that node has same attributes then it considered to be linked [2]. Forth approach of link prediction is based on the machine learning approach using supervised learning. This approach used machine learning algorithms like logistic regression, decision tree, naïve bayes or random forest to find out potential links between nodes on number of features extracted from network. Machine learning algorithm has potential to iterate over each attribute and find out the similarities between the nodes. Machine learning algorithm get trained on the extracted data of network and if any new node come it find out that this new node belongs to any existing node from the list of nodes. This approach is useful as compared to other approaches because machine learning algorithm can easily understand the complex structures and large network while other approaches cannot handle

the complex structure of nodes. Where this approach is useful, but it also requires a huge amount of get train on the network [3].

- **Link Prediction Applications.** Link prediction can be used in various fields including social network analysis, marketing, recommendation systems, fraud detection. Link prediction can be used in social network analysis by identifying the relationship between nodes in a network. It can be helpful to identifying the communities to spread information. In marketing link prediction can be used as a tool to find out the potential customers based on their social connections. In recommendation systems links prediction can be used to recommend or suggest new products based on the past behavior of customers. In fraud detection link prediction can be very useful it can find out the nodes that had done fraud in past and based on the attributes of those fraudsters nodes if any new node appears with the attributes of fraudsters node it will highlight that node [4].

Approach:

To perform the link prediction on reddit hyperlink network dataset I used the logistic regression model which is supervised learning algorithm which take the input data with the labels to predict the outcome. Logistic regression algorithm has two output 0 or 1. To predict the hyperlink between nodes logistic regression is used if the output of the logistic regression model is 1 it means it exists as relationship between two nodes and if the output is 0 it means there is no relationship between two nodes.

Below are the steps to perform the link prediction using logistic regression:

- **Data Preparation.** Before training the logistic regression the dataset of hyperlink was prepared. All the data processing techniques are applied before training logistic regression. The features I picked for training logistic regression is number of common neighbors between two subreddits, similarity of their node attributes, distance of nodes in the network. Dataset is divided into the training and the testing phase.
- **Data Exploration.** After preparing the data the data exploration is carried out to understand the distribution of the data.

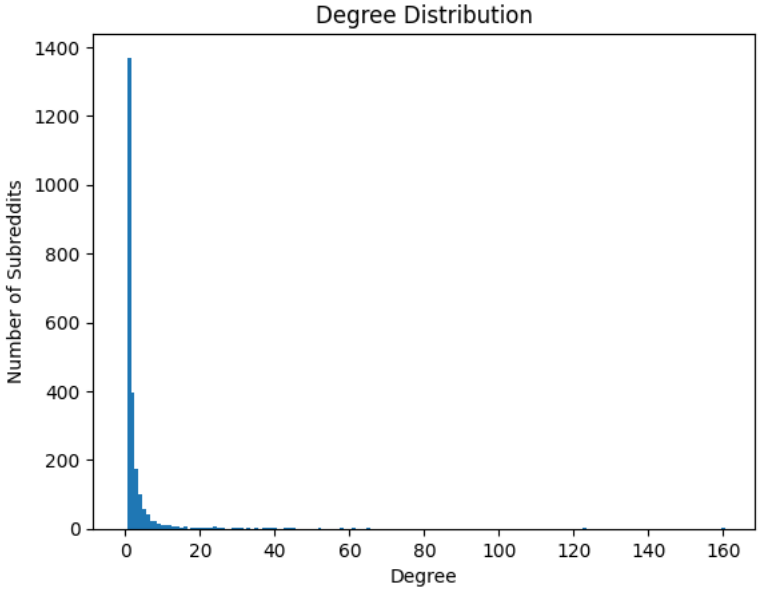

Figure 1. Degree of Distribution

In the above figure 1 the visualization is created to show the distribution of number of connections each node has in the network. The above visualization gives overall structure of the network and how well-connected individual subreddits are.

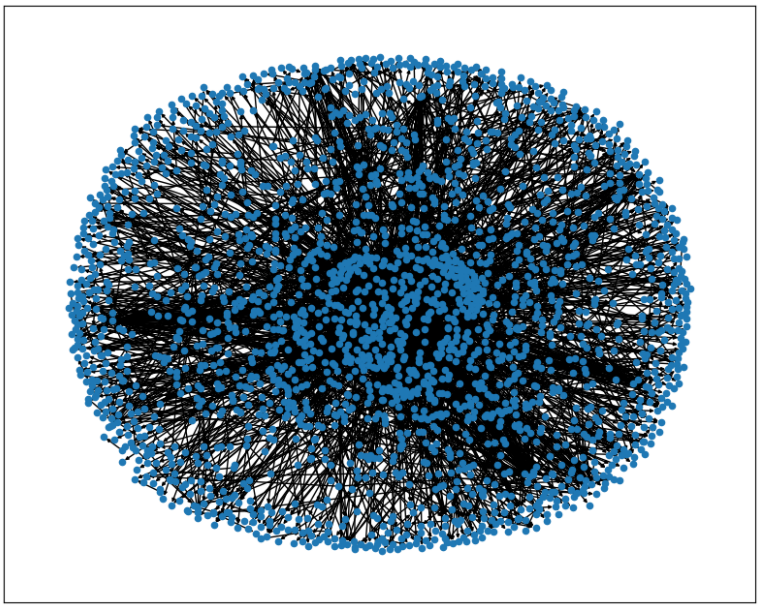

Figure 2. Spring Layout of graph

In the figure 2. The spring layout of the graph is plotted.

- **Model Training.** Training data is used for training the logistic regression to predict the presence of hyperlink between two subreddits based on the input features. In training logistic regression learn from the input data samples and prepare the model for making predictions. For training the logistic regression model the scikit learn library is used.
- **Model Evaluation.** After training the logistic regression model the model will be evaluated using different evaluation metrics on testing dataset. F1-score, recall and precision are used to evaluate the logistic regression model. If the model performs well on these evaluation metrics shows good results, it means the model has good training over the training data.
- **Link Prediction.** After training and the evaluation of model the link will be predicted using the input features. Input features including nodes, their adjacent nodes, distance with the adjacent nodes and the result of the model for these inputs will be is there exist a link between the node or not.

Experiment:

For performing the link prediction on the dataset of the reddit hyperlink network logistic regression is used. The dataset was divided into two parts training and testing. Features used for training logistic regression model was number of common neighbors between two subreddits, similarity of node with other nodes, distances of nodes in the network. The output variable is presence of a hyperlink between two subreddits. Logistic Regression model was trained on the scikit learn library. Once the model gets trained it was evaluated on the different evaluation metrics including F1 score, Precision and Recall. The model shows promising evaluation results the accuracy of the model was 1.0, f1-score of models was 1.0, recall of the model was 1.0 and recall of the model was 1.0. Logistic Regression is used to predict the presence of the hyperlink between two nodes based on the input features. Input features the nodes, adjacent nodes and the distance of nodes with their adjacent nodes. Results of the logistic regression algorithm can be used effectively to perform link prediction on the reddit hyperlink network dataset. Effective features were selected for training of the logistic regression to predict the hyperlink. High accuracy, precision, f1-score and recall values of the logistic regression model states that this model can be used for further analysis and prediction on the dataset.

Conclusion:

This research was focused on the prediction of using the reddit hyperlink network dataset. To identify the links between two nodes in the network and to gain the insights of the structure of the nodes was the aim

of this research. Dataset is ranging from January 2014 to April 2017 the dataset contains 55863 nodes and total 858,490 edges. For performing link prediction, the useful input features were passed to the logistic regression to train the model and predict the link between two nodes. Logistic regression model was evaluated on different evaluation metrics, model has shown the promising results. Results of this research can be used for many practical applications, including social network analysis, recommendation systems, fraud detection and marketing. By carrying this research in more depth more insights can be generated like how people connect and interact with each other online and these insights can be used more effectively for communication and for marketing purposes.

REFERENCES

1. Daud, Nur Nasuha, et al. "Applications of link prediction in social networks: A review." *Journal of Network and Computer Applications* 166 (2020): 102716.
2. Wang, Peng, et al. "Link prediction in social networks: the state-of-the-art." *arXiv preprint arXiv:1411.5118* (2014).
3. Govinda, K., et al. "Link Prediction in Social Networks using Machine Learning." *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. IEEE, 2023.
4. Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F., & Anuar, N. B. (2020). Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications*, *166*, 102716.