Social Network Analysis for Precise Friend Suggestion for Twitter by Graph Embedding Using Node2vec

Submitted 6 November 2023, in partial fulfillment of
the conditions for the award of the degree **BSc in Internet of Things and Robotics Engineering.**

Md.Shahriar Hossain Apu ,Mehedi Hasan Emon,Mahjabin Ashka,Arifuzzaman Ripon
1901036.1901031,1901048,1901038

Supervised by Mahir Mahbub

Bangabandhu Sheikh Mujibur Rahman Digital University

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature ____________________

Date ______ / ______ / ______

Abstract

The main aim of this paper is to create a friend suggestion algorithm that can recommend new friends to a user on Twitter based on their existing friends and other relevant details. The information used for making these predictions includes the user's friends, tags, tweets, language spoken, ID, and more. The authors employed supervised learning methods to train their models, specifically using the Node2Vec approach. Node2Vec is a machine learning technique used to learn embeddings of nodes in a network, which can capture the structural characteristics of the social graph. By using Node2Vec, the authors sought to capture the nuanced relationships between users and their friends.The approach is user-centric, focusing on a single user at a time. For each user, their features are compared to those of other users, particularly their non-friends, to suggest potential new friends. This approach takes into account the uniqueness of a user's social network and community, making personalized friend recommendations based on the learned node embeddings.

Acknowledgements

We would like to express my heartfelt gratitude to my supervisor, Mahir Mahbub, for his invaluable guidance and unwavering support throughout the course of this research. His expertise, mentorship, and insightful feedback were instrumental in shaping the direction of this project. We are truly fortunate to have had the opportunity to work under his supervision, and We are deeply appreciative of his dedication to my academic and research pursuits.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The paper discusses Twitter, a popular microblogging platform with a history of approximately 13 years and a continually growing user base. On Twitter, users are provided a platform to express their thoughts and opinions, limited to 140 characters. Users can follow other users to receive their tweets, leading to the formation of communities with shared interests. Node2Vec is employed to capture the structural characteristics of the social network by learning embeddings of nodes in the network. This technique helps in identifying users with similar interests and connections. In contrast to the[3] K-Nearest Neighbors (KNN) and [2]Word2vec approach, which uses distance-based similarity, Node2Vec utilizes network-based similarity to make recommendations. [1]Node2Vec is utilized to compare the features of the central user's non-friends, which enables the identification of suitable friend suggestions. Given that each user's friends and communities are unique, the research emphasizes the need to focus on individual users. For this purpose, a single user is selected as the "central user," and the results obtained for this user are used to generate friend recommendations. This user-centric approach is fundamental to the research, as it recognizes the distinctive nature of each user's social network. Therefore, the model is effectively trained based on the characteristics and preferences of the central user. The paper contributes to the field by addressing the challenges and uniqueness of friend recommendation within the context of Twitter's dynamic and diverse user base, using the Node2Vec approach to capture the underlying network structure.

1.1 Motivation

Motivation for the research presented in this paper is rooted in the evolving landscape of social media, with a specific focus on Twitter, a popular microblogging platform. Here are the key motivators for this study:

1. Twitter's Growing User Base: Twitter has amassed a substantial and continually growing user base over its approximately 11-year history. As more users join the platform, there is an increasing need for effective and personalized services to enhance their experience.

2. User Engagement and Community Building: Twitter users actively engage in discussions, sharing their daily activities, opinions on current events, and interests. This platform serves as a hub for users with similar intentions to connect and form communities based on shared interests.

3. Friendship and Social Network Dynamics: In the realm of social media, forming and nurturing friendships is a fundamental aspect of user engagement. Understanding how friendships are forged and maintained on platforms like Twitter is pivotal to user satisfaction and retention.

4. Personalized User Experience: As Twitter's user base diversifies, the need for personalized experiences becomes more pronounced. Users seek relevant content and connections tailored to their interests, and a robust friend suggestion algorithm can play a pivotal role in achieving this.

5. Unsupervised Learning and Node2Vec: The choice of utilizing supervised learning methods, particularly the Node2Vec approach, presents an opportunity to delve into the nuances of Twitter's social network. The research leverages this advanced machine learning technique to capture the structural characteristics of the network and provide personalized friend recommendations.

6. User-Centric Approach: The paper recognizes the uniqueness of each user's social network and emphasizes a user-centric approach to friend recommendations. Understanding that each user's friends and communities are distinct, the research tailors its approach to individual users, selecting a central user as a reference point.

7. Challenges in Friend Recommendation: The challenges and intricacies of friend recommendation on a dynamic and diverse platform like Twitter provide fertile ground for research. The paper seeks to address these challenges and offer valuable insights into the development of effective friend suggestion algorithms.

1.2 Aims and Objectives

In this paper, the primary objective is to develop a friend suggestion algorithm tailored to Twitter. This algorithm aims to recommend new friends to a user based on the user's existing connections and additional user details. The predictive factors include information such as the user's friends, tags, tweets, language preferences, and user ID. The models are trained using unsupervised learning methods, with a specific focus on the Node2Vec approach.

Aims:

1. Enhance Twitter User Experience: The primary aim of this research is to enhance the user experience on Twitter by developing a robust friend suggestion algorithm that provides users with relevant and personalized friend recommendations.

2. Improve Engagement and Community Building: The research aims to foster increased user engagement and community building on Twitter by facilitating meaningful connections among users with similar interests and activities.

3. Utilize Advanced Machine Learning Techniques: This research seeks to harness the potential of advanced machine learning techniques, particularly Node2Vec, to capture the structural characteristics of Twitter's social network and use them for improving friend recommendations.

4. User-Centric Approach: An essential aim is to adopt a user-centric approach that acknowledges the uniqueness of each user's social network and focuses on individualized friend suggestions, thereby delivering a more tailored and engaging experience.

5.

Objectives:

1. Data Collection and Preprocessing: Collect and preprocess data, including user information, friends, tags, tweets, language preferences, and other relevant data, to create a foundation for the friend suggestion algorithm.

2. Unsupervised Learning Model Development: Employ supervised learning methods to develop a predictive model that can learn from the user data and generate accurate friend recommendations.

3. Node2Vec Integration: Integrate the Node2Vec approach into the model to extract structural insights from the Twitter social network, allowing for network-based similarity in friend recommendations.

4. Dimensionality Reduction: Utilize Node2Vec to reduce the dimensionality of the data and feature space, enabling more efficient and effective friend suggestions.

5. User-Centric Friend Recommendations: Implement a user-centric approach by selecting a central user as the focal point and comparing their features to those of non-friends to identify suitable friend suggestions.

6. Evaluation and Validation: Assess the performance and effectiveness of the friend suggestion algorithm through rigorous evaluation and validation processes, ensuring that it provides meaningful and accurate recommendations to users.

7. Contributions to Twitter's User Experience: Contribute to Twitter's commitment to delivering a user-centric experience by providing an improved friend suggestion feature that enhances user engagement and community building.

8. Research Dissemination: Share the findings and insights gained from this research through academic publications and presentations, contributing to the broader field of social network analysis and recommendation systems.

1.3 Description of the work

In this proposed work, used the Twitter API to collect the user data in the following way. Data collection steps taken can be listed as follows: choosing the central user, find their set of current friends, finding the set of users that are not friends with the current user (non-friends), and finally, gathering the profile information for every friend and non-friends concerning the central user. Now coming to all the major parts of the work in detail.

1.3.1 Dataset

We collected the dataset from[4]kaggle.Twitter enables us to mine the information of any client by utilizing Twitter API or Tweepy libraries in Python. The information extracted will be the tweets done by the users in the given time, along with their other details. The primary activity here involves getting the buyer key. These keys will help the API for confirmation. Data collection was hampered by the API's restrictions and privacy settings. To begin with, some profiles only allow limited access, making it impossible to view any information other than the user's name in certain cases. As a consequence, in this work, limited ourselves to selecting profiles inside particular networks that supplied us with all of the information, the author need, making it impossible to ask for all of a user's friends directly. The dataset can only be used to identify whether or not two people are friends. As a consequence, in this work, needed a large number of user ids to begin with, and it was impossible to find all of a person's friends unless had a large enough starting set. The lack of a straightforward mechanism for creating a list of users based on whatever arbitrary criteria which selected was the issue to be considered.

1.3.2 Choosing Central User

Since suggesting friends for each user would require querying the dataset for every user again and again. This implies that the algorithm would have a very high time complexity to be able to run in real-time. Due to this issue, in this work, fix a central user for which all the quantitative analysis is done once, and the corresponding friends are suggested. This central user can be changed every time the algorithm is tested. Once the algorithm is tested for the central user, it can further be extended to work for all the users independently in a parallel computing platform.

1.3.3 Comparison of Parameters Among Non-Friends

Once the dataset was queried and here in this paper, chose the central user to start with, the parameters obtained in the dataset were compared to decide on the possible factors that can affect the friend suggestion algorithm. While the number of followers can be a factor in determining the possible friends of a user, it is already incorporated by Twitter so other factors like the tag of the tweet, tag popularity, and common friends were explored further. Also, the subject of the user's tweet was obtained by using natural language processing and then compared with tags of other users' tweets to further enhance the suggestions.

1.3.4 Generating Friends List

Figure 1 shows the flow diagram for Friend Suggestion in Twitter. For this purpose, the non-friends of users are ranked as per the number of friends they have in common with the central user. Apart from this, the overall count of the tag of a tweet posted by the central user is compared with all other tags to evaluate the popularity of the user's tweet's tag. The tweet posted by the user is then analyzed for its subject and this subject obtained from the text is then compared with tags of other user's tweets

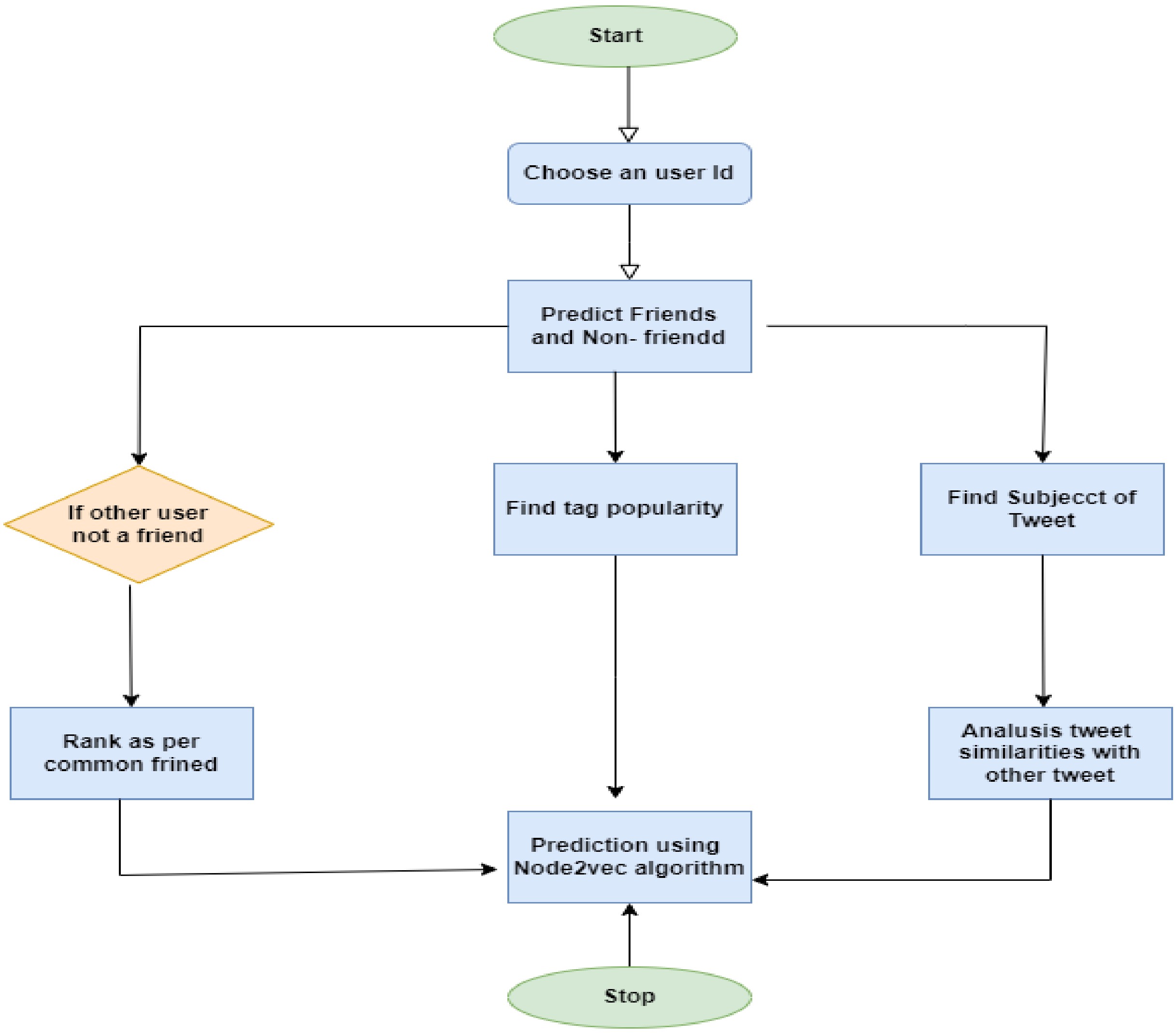

Figure 1.1: Flow Diagram of Friend Suggestion Algorithm

```
node2vecWalk (Graph G' = (V, E, π), Start node u, Length l):

    Initialize walk to [u]
    for (walk_iter = 1 to l):
        curr = walk[-1]
        V_curr = GetNeighbors(curr, G')
        s = AliasSample(V_curr, π)
        Append s to walk
    return walk
```

```
Learn_Features (Graph G = (V, E, W)):

    Dimensions -> d
    Walks per node -> r
    Walk length -> l
    Context size -> k
    Return -> p
    In-out -> q

    π = PreprocessModifiedWeights    (G, p, q)
    G' = (V, E, π)

    Initialize walks = 0

    for (iter = 1 to r):
        for (all nodes u ∈ V):
            walk =     node2vecWalk(G', u, l)
            Append walk to walks

    f = StochasticGradientDescent(k, d, walks)
    return f
```

Figure 1.2: Node2vec Algorithm

Chapter 2

Background and Related Work

[6]The author talks about different networks and finds the relationship to recommend friends. It has two major components first being related networks by selecting important features and the second being network structure and preserving most of it. It is based on friend correlation and considers effect in different social roles. Huang et al. (2016). The author gives idea about a new friend recommendation system using artificial bee colony(ABC) which indicates a link between users. [7]It is based on the structural properties of the social network. Firstly, it finds the relevant parameters for the relationship among users using social topology. The sub- graph of the network is composed of users and all users within the network separated by three degrees of separation, then based on the subgraph new links are suggested thus indicating new friends. Akbari et al. (2013). [9]The paper gives a idea about a more precise friend recommendation with 2 stages. In the first stage the information of the relationship between text and users, then align the recommended friends. In the second stage, they built the topic model of the relationship between image features and users. Huang et al. (2017). [8]The paper proposed a novel semantic-based recommendation of friends based on their lifestyles. They take advantage of smartphones, friend books to discover the lifestyle of a user from user-centric data and then measure the similarity of lifestyle to finally recommend from with similar lifestyles. The friend book finally keeps a list of people with the highest score to recommend a friend. Wang et al. (2015). The paper gives idea about users who want to meet friends on social media, they interviewed active users and then developed a friend request acceptance

model to refer to various factors that influence it. They found out the major factor that impacts the person who accepts the friend request, mostly person with common hobbies and mutual friend is accepted. Rashtian et al. (2014). The paper talks about how to find short paths between users in a network which would indicate [5]their closeness and will also result in them being a good recommendation as a friend. These are based on email contact, it is found that the kind of things people talk or find is a huge factor in determining their closeness. Adamic and Adar (2005). The paper "Ranking Users for Intelligent Message Addressing" uses various Machine Learning algorithms for its task. It also works on the simple k nearest neighbors concept to overtake other algorithms. It finds people who can receive a message from the central user, the person who is looking for friends. It understudies various people to conclude that intelligent messages can be added advantage to email Carvalho and Cohen (2008). The paper "Inferring relevant social networks from interpersonal communication" analyzes the network to look for the region of interest. It sees large social networks to look for unobserved ties (the tie in which i and j are connected) in the event (where i email j). The paper had two conclusions first being prediction task choosing threshold value yields better results and secondly, optimal threshold value seems to be consistent.

Chapter 3

Design

3.1 SYSTEM ARCHITECTURE

The objective of this section is to discuss a novel system architecture for friend suggestions in a Twitter- like framework. This is done by working with the proposed methodology discussed in the previous section. The corresponding system architecture is illustrated in Figure 3.1

While designing the system, in the first phase, that is, the data mining phase, the dataset is extracted from Twitter. In the analysis and computation phase, one part deals with the data analysis while another module deals with the tweet analysis part of the system. The component of the system that analyzes the connection among users is included in the data analysis part. As shown in the figure, the system architecture has four major independent modules, which are described in detail.

3.1.1 Data Extraction Module

In this module, libraries such as Tweepy are used to extract user data from Twitter. Due to security reasons, only a few required details for each user are extracted. Also, the user-selected belong to a common social group so that their friends are also present in the dataset. Once data is extracted, the friends of the users that are not a part of the dataset are excluded during the preprocessing stage. The output of this component is the

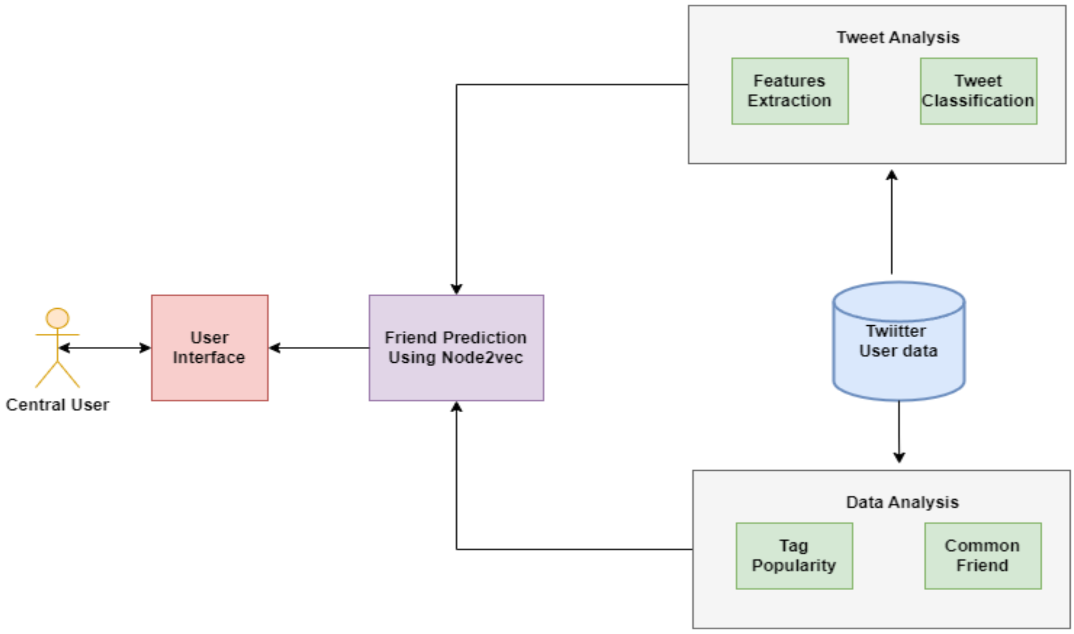

Figure 3.1: System architecture for friend suggestion on Twitter

processed CSV file containing data for each user.

3.1.2 Data Analysis Module

Given the tags of tweets available for each of the users, this module does important computations for detecting similarities among various non-friends of the central user. This module further has two parts, that is, calculating tag popularity and finding common friends. In this phase, the mathematical computations take place and all the quantitative results are obtained. Other smaller attributes that are considered for calculations are the language spoken by the current user, their number of followers, etc. These attributes are already a part of the Twitter friend suggestion algorithm. However, they also play a role in detecting similarities among two users.

3.1.3 Tweet Analysis Module

This component deals with extracting the subject of the tweet from the content of the tweet. This subject is then compared with the tags of all the other tweets to analyze the similarity between them. This is done by using chunking, then feature extraction is done on the obtained chunks of sentences in the tweet. Once this is done, the features and the probability of that feature being the subject of the tweet are compared to finally arrive at the resultant subject.

3.1.4 Similarity Detection Module

Once all the results from the data analysis component and the tweet analysis component of the system are obtained, these results are then further combined to arrive at the final similarity metric for the non- friend. This is done by calculating the distances between the data for the given central user and the current user.

Chapter 4

Implementation

4.1 IMPLEMENTATION

4.1.1 Description of Dataset

The dataset used is directly extracted from Twitter using the Tweepy library in Python. It consists of a network of 4000 users. The attributes included in the dataset are username, user id, user avatar, tweet, tweet id, tweet tag, language, number of followers, and id of all the friends of the user. The dataset description is discussed in Table 1. The table shows the overall analysis of the data extracted from Twitter. Figure 3 illustrates the graphical distribution of various languages used by all the users in their tweets. The Figure shows that most of the users in the dataset communicate in English

id	screenName	tags	avatar	followersCount	friendsCount	lang	lastSeen	
"1969527638"	"LIngoMakeEmCum_"	["#nationaldogday"]	"http://pbs.twimg.com/profile_images/534286217...	319	112	"en"	1472271687519	"7693107
"51878493"	"_notmichelle"	["#nationaldogday"]	"http://pbs.twimg.com/profile_images/761977602...	275	115	"en"	1472270622663	"7693094
"1393409100"	"jesseayye"	["#narcos"]	"http://pbs.twimg.com/profile_images/713282938...	120	107	"en"	1472804144409	"7716226
"232891415"	"MrBrianLloyd"	["#gloryoutnow"]	"http://pbs.twimg.com/profile_images/133440668...	492	325	"en"	1472269186776	"7693081
0422907207680"	"sarahdorat_16"	["#nationaldogday"]	"http://pbs.twimg.com/profile_images/767180520...	128	218	"en"	1472271397356	"7693097

Figure 4.1: Snapshot of Dataset

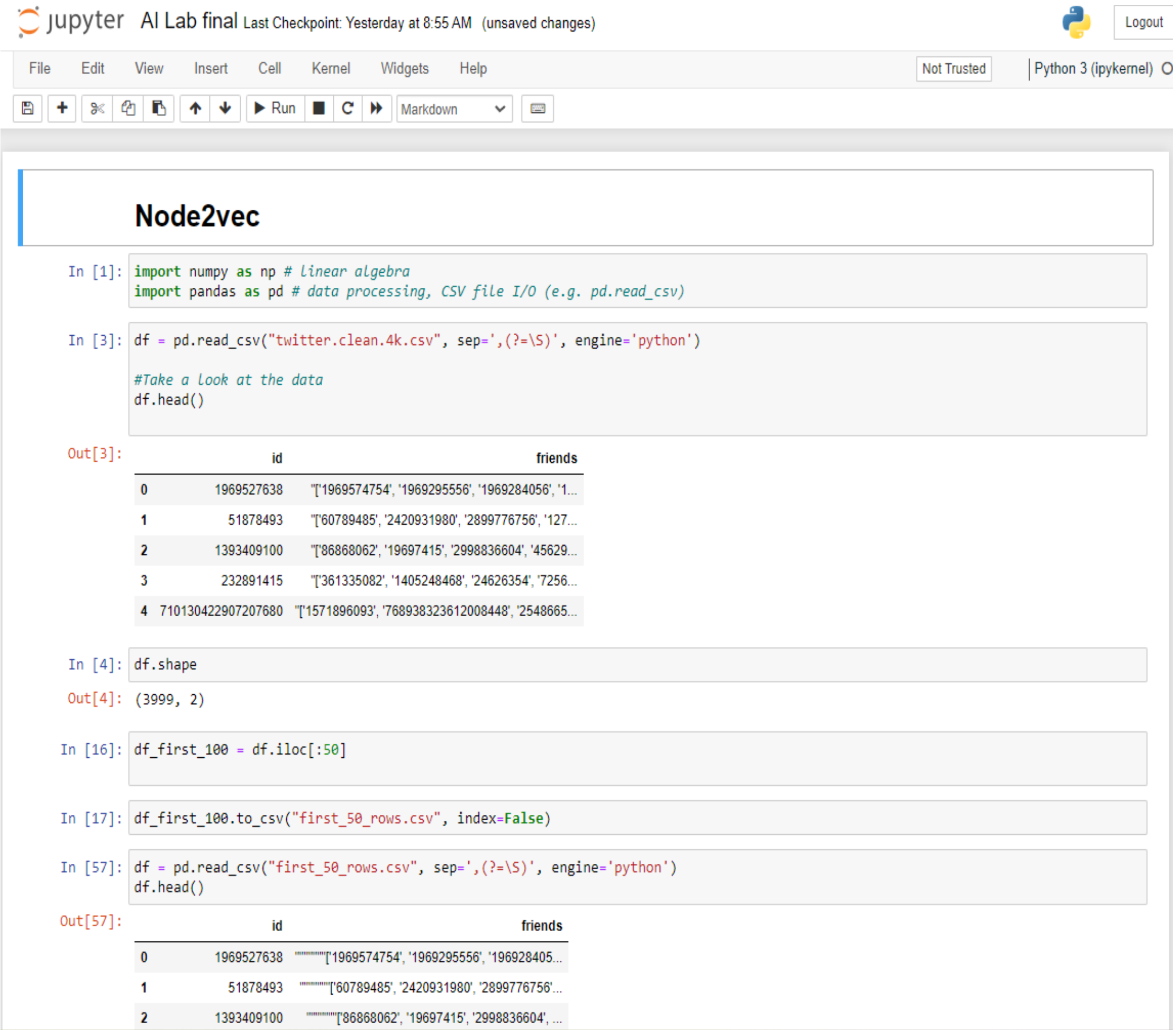

Figure 4.2: Working Environment

4.1.2 EXPERIMENTAL SETUP

The execution of the code required the installation of Python 3 on the system using Jupyter showing in figure 4.2- Along with this, feature extraction and subject determination of tweet require NLTK module in Python, which enables natural language processing in Python. The central user is randomly selected and given as input while running the code.

Chapter 5

Evaluation

5.1 RESULTS AND ANALYSIS

The dataset used here has a network of 4000 users, where all the friends of the user are in the network itself and even if a user has friends that are not a part of the dataset, they are excluded from the

Table 5.1: Dataset

Description Number of users	4000
Number of distinct tags	483
Avg. number of followers per user	390 Avg. number of friends per user

analysis part. The algorithm suggests the top 10 friend suggestions that a user can relate

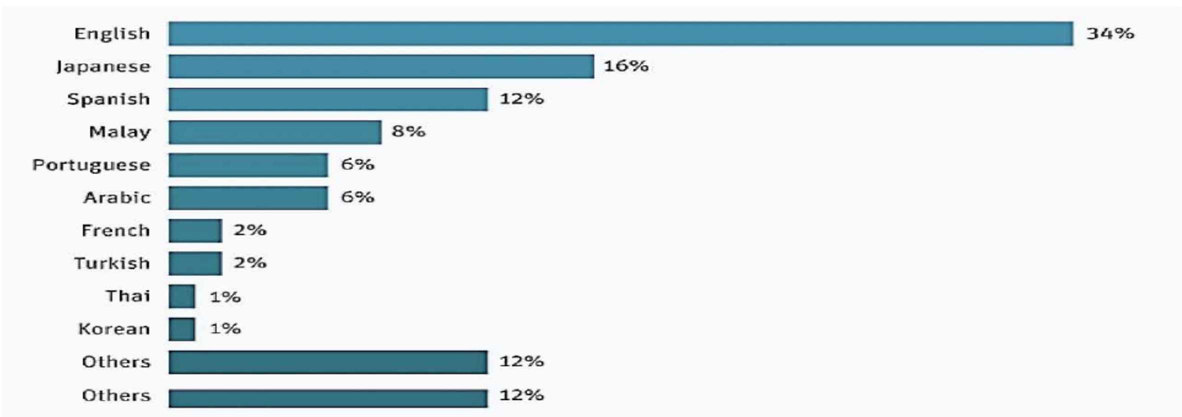

Figure 5.1: Language distribution in the dataset

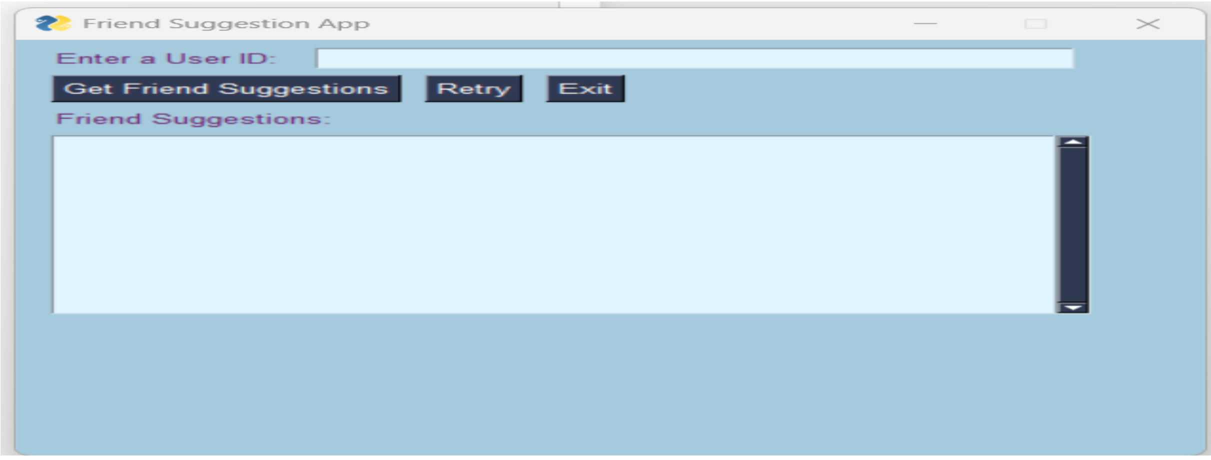

Figure 5.2: User Interface of Frined Prediction

Figure 5.3: Suggested Friend in the Window

with as per the quantitative results obtained from the factors studied. However, if there is a clash in a ranking of several users, that is, if multiple non-friends have the same final score, then the users are displayed in alphabetical order. This means that there is a chance that a user who may be a possible friend of the central user is excluded from the suggestions because of the clash in the rank. Figure 5.2 is the user interface and Figure 5.3 show the top friend suggestions for two different users having the same tags. Here, it can show that have similar friend suggestions as they posted tweets with the same tags. Although the results that have been discussed above were obtained by examining the dataset ex- tracked from Twitter, the approach and the algorithms described in the paper

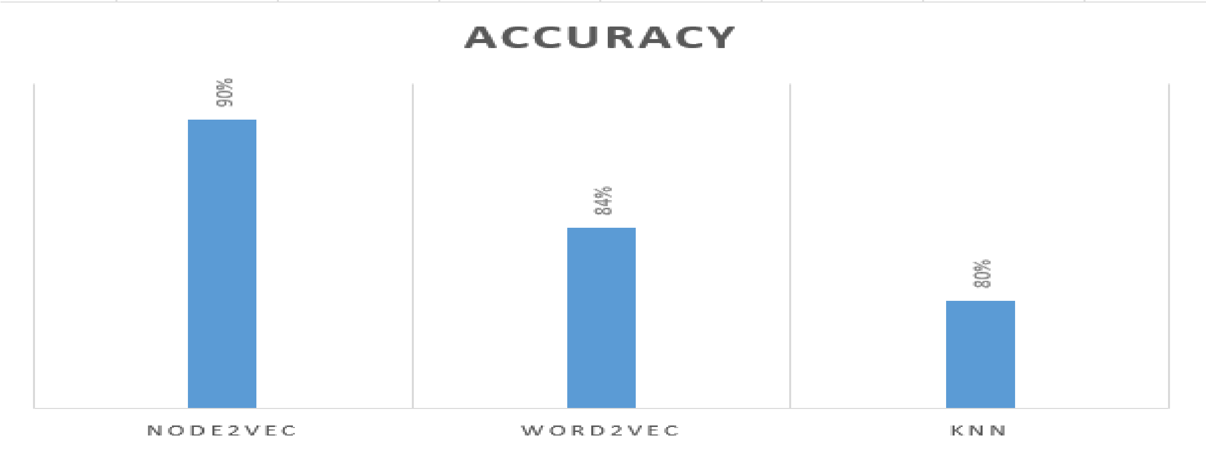

Figure 5.4: Model Performance

can be applied to any social networking based framework, such as Facebook or even e-mail based framework, such as Gmail (where it can be used to suggest email groups).

Table 5.2: Model Accuracy

Model Name	**Accuracy %**
Node2vec	90%
Word2Vec	84%
KNN	80%

Chapter 6

Summary and Reflections

6.1 CONCLUSION

The idea behind our friend suggestion method is that people have a greater affinity for their friends than for strangers. As a result, non-friends who have a common friend have greater clout than nonfriends who are disconnected from the core user. Having a high number of common friends does not necessarily mean the user's interests are aligned with the central user's. Non-friends must be ranked according to a set of criteria that takes into consideration the users' common interests. This would help to separate those with different interests from those who have similar ones, boosting the algorithm's efficiency. In this research work, proposed a friend recommendation algorithm based on Tweet similarity and Tag popularity, in addition to the existing criterion of mutual friends. Nonfriends may be ranked based on these factors by our recommendation engine. Using natural language processing, provided techniques for assessing tag popularity, finding common friends, and measuring tweet similarity. Our technique may improve Twitter's existing friend suggestion algorithm by offering a more efficient and complete approach to friend recommendation, according to the study of the small network dataset gathered. In simple words, instead of recommending friends based on common friends, our algorithm takes into account extra platform-relevant factors. Finally, in this research contribution, have tried to improve the problem of identifying a user's offline (that is, real-life) social community, purely from examining the Twitter network structure of the central user.

Based on observations from our Twitter data and results from previous works, it propose an algorithm involving three factors to improve the Twitter friend suggestion algorithm. Incorporating these factors, developed a novel algorithm to iteratively discover the user's possible friends based on a new way of measuring user closeness.

6.2 FUTURE WORKS

The current[5] algorithm being used in the friend suggestion is the K-nearest neighbors algorithm. While this algorithm helps derive useful interpretations from the dataset, comparing quantitative results from other algorithms can help improve the suggestions further. Other training and classification algorithms such as Convolutional Neural Networks and Support Vector Machines can be used and differences in accuracy rate can be analyzed and compared among the various algorithms to arrive at a model that gives the best results for friend recommendation in Twitter. Other factors such as the popularity of user's tweets can also be analyzed to make interpretations and use them to make suggestions better. There can also be amendments that can be made to the algorithm to make it faster and more time- efficient. Another very meaningful similarity metric that can be used to derive results is cosine similaritybased comparison. It measures the similarity between two non-zero vectors by measuring the cosine of the angle between them. The lesser the angle, the more similar/related are the corresponding users.

Bibliography

[1] https://www.geeksforgeeks.org/node2vec-algorithm/.

[2] https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/.

[3] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

[4] https://www.kaggle.com/datasets/hwassner/twitterfriends.

[5] Akbari, F., T. A. H. . N. A. F. Graph-based friend recommendation in social networks using artificial bee colony. *In Dependable, Autonomic and Secure Computing. DASC. 27*, 3 (2013), 379–423.

[6] Huang, S., Z. J. W. L. . H. X. S. Social friend recommendation based on multiple network correlation. *IEEE Transactions on Multimedia 18*, 2 (2016), 287–299.

[7] Java, A., S. X. F. T. . T. B. Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*.

[8] Parveen, R., and Varma, N. S. Friend's recommendation on social media using different algorithms of machine learning. *Global Transitions Proceedings 2*, 2 (2021), 273–281. International Conference on Computing System and its Applications (ICCSA- 2021).

[9] Singh, D. K., N. N. R. L. S. P. V. R. S. M. P. H. M. . T. G. B. Social network analysis for precise friend suggestion for twitter by associating multiple networks using ml. *International Journal of Information Technology and Web Engineering (IJITWE)*, 17, 1–11.

Appendix A

User Manuals

The user have to to input a user id in the highlighted bar showed in figure A.1 - and the output will show like this figure A.2:

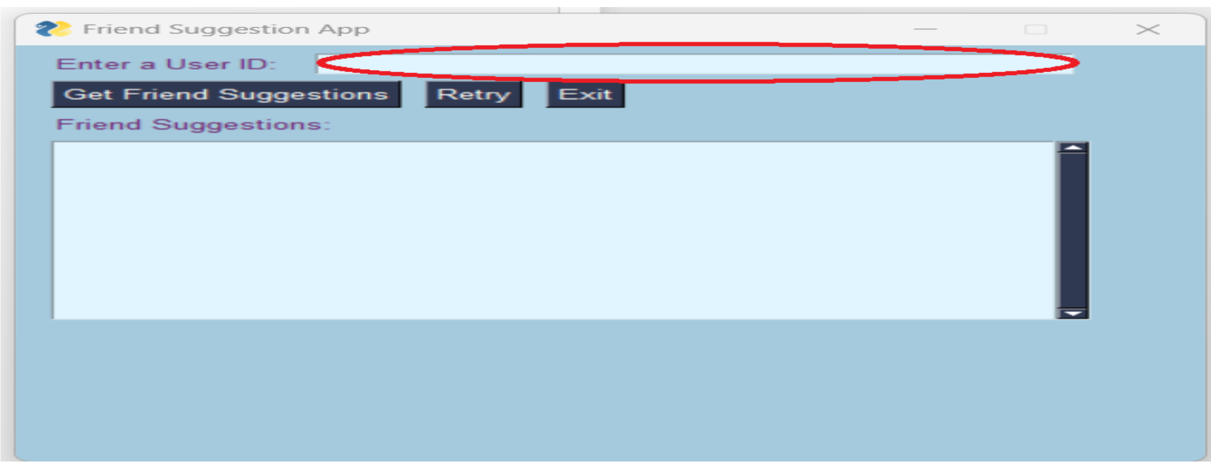

Figure A.1: User Input

Figure A.2: Output