Using Homophily to Analyze and Develop Link Prediction Models with Deep Learning Framework

by

Kazi Zainab Khanam

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

in

Computer Science

in the

Faculty of Science and Environmental Studies

of

Lakehead University, Thunder Bay

Committee in charge:

Dr. Vijay Mago (Principal Supervisor)

Dr. Rajesh Sharma (External Examiner)
Dr. Yimin Yang (Internal Examiner)

Winter 2021

The thesis of Kazi Zainab Khanam, titled Using Homophily to Analyze and Develop Link Prediction Models with Deep Learning Framework, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

Lakehead University, Thunder Bay

Using Homophily to Analyze and Develop Link Prediction Models with Deep Learning Framework

Kazi Zainab Khanam

Abstract

USING HOMOPHILY TO ANALYZE AND DEVELOP LINK PREDICTION MODELS WITH DEEP LEARNING FRAMEWORK

Twitter is a prominent social networking platform where users' short messages or "tweets" are often used for analysis. However, there has not been much attention paid to mining the medical professions, such as detecting users' occupations from their biographical content. Mining such information can be useful to build recommender systems for cost-effective advertisements. Conventional classifiers can be used to predict medical occupations, but they tend to perform poorly as there are a variety of occupations. As a result, the main focus of the research is to use various deep learning techniques to examine the textual properties of Twitter users' biographic contents, network properties, and the impact of homophily of Twitter users employed in medical professional fields. In Chapter 2, a survey is presented based on the concept of homophily as well as important social network topics that summarize the state of art methods that has been proposed in the past years to identify and measure the effect of homophily in multiple types of social networks. This enables us to find open challenges and directions for future research. In Chapter 3, a model has been developed to identify Twitter users working in medical professional fields by using textual properties of the Twitter Users' bio contents. We have conducted our analysis by annotating the content of Twitter users' bios and propose a method of combining word embedding with state-of-art neural network models. Finally, in Chapter 4, the research introduces a link prediction model based on the homophily concept by using the Twitter users' followers and following IDs identified from Chapter 3. Recent research has centered on analyzing rapidly

evolving networks. While predicting links in dynamic networks is difficult, deep learning techniques and network representation learning algorithms, such as Node2vec, have demonstrated significant improvements in prediction accuracy. However, Node2vec's Stochastic Gradient Descent (SGD) approach is prone to falling into a local optimum, and as a consequence, Node2vec fails to capture the network's global structure. To address this problem, we propose NODDLE (integration of NOde2vec anD Deep Learning mEthod), a deep learning system in which we combine Node2vec's features and feed them into a four-layer hidden neural network. integration of NOde2vec anD Deep Learning mEthod (NODDLE) takes advantage of adaptive learning optimizers for improving the performance of link prediction. On different social network datasets, experimental findings show that our approach outperforms conventional methods.

Dedication

This is dedicated to my family and friends for their unwavering support, as well as the educators who believed in me during my academic career.

Contents

List of Figures

List of Tables

Acronyms

AA Adamic Adar. 63

ADADELTA An Adaptive Learning Rate Method. 63

ADAGRAD Adaptive Gradient Algorithm. 63

ADAM Adaptive Moment Estimation. 63

ALBERT A lite BERT. 42

BERT Bidirectional Encoder Representations from Transformers. 42, 45, 47, 52, 58

BiLSTM Bidirectional LSTM. 42, 46, 55

GRU Gated Recurrent Unit. 42, 44, 46, 52, 55

JC Jaccard co-efficient. 63

LSTM Long Short Term Memory. 42, 44, 46, 52, 55, 58

NOC National Occupational Classification. 50, 51

NODDLE integration of NOde2vec anD Deep Learning mEthod. 2, 3

PR Preferential Attachment. 63

SGD Stochastic Gradient Descent. 58, 63

VSM Vector Space Model. 42, 47, 52

Acknowledgments

I would also like to express my gratitude to my supervisors: Dr. Vijay Mago and Dr. Gautam Srivashtava and my colleagues at Lakehead University's DaTALab in the CASES building for their invaluable knowledge and experience. I would like to express my gratitude to Dr. Abhijit Rao for proofreading the work in Chapter 2. Andrew Heppner and Maegen Lavallee also deserve special thanks for their help with scholarship applications and for addressing my numerous questions. I am grateful for the grant from the Natural Sciences and Engineering Research Council (NSERC), as well as the resources provided by the DaTALab, for supporting my work during my degree. In addition, an NSERC Discovery Grant provided by my supervisor, Dr. Vijay Mago, covered the costs of publication and other expenses.

I would like to thank DaTALab members & Lakehead University's HPC (High Configuration GPU enabled PC) for executing the models, and Mohiuddin Md Abdul Qudar Punardeep Sikkha, Arunim Garg, Bart, and Abhijit Rao for proofreading and reviewing the manuscript. This research is funded by NSERC Discovery Grant (RGPIN-2017-05377), held by Dr. Vijay Mago.

Chapter 1

Introduction

The main focus of this thesis is to analyze the textual properties of Twitter users' biographic contents, network properties, and the effect of homophily in medical professional fields by using various deep learning techniques. In Chapter 2, the thesis provides a background on the concept of homophily related to social networks. It also acts as a resource for researchers to gain an understanding of the state-of-the-art models for analyzing the impact of homophily on social networks.

In Chapter 3, this thesis presents a model to identify Twitter users working in medical professions based on the textual properties of the Twitter users' bio contents. Twitter is a popular social networking site, and user's post or 'tweets' have been used extensively for research purposes. However, not much research has been done in mining the medical professions, such as determining the occupations of users from their biographical contents. Mining such information can be useful for building efficient recommender systems for cost-effective advertisements. For example it is more effective to advertise the journals of the latest medical articles to a network of doctors than to users who are working in IT sectors. Moreover, it is highly important to develop effective methods to identify the occupation of users since con-

ventional classification methods rely on features developed by human intelligence. Although the result may be favorable for the classification problem, it is still extremely challenging for traditional classifiers to predict medical occupations accurately since it involves predicting multiple occupations. Hence this study emphasizes on predicting medical occupational class of users through their public biographical ("Bio") content. Our analysis is conducted by annotating the bio content of Twitter users. A method is proposed that combines word embedding with state-of-art neural network models that include: Long Short Term Memory (LSTM), Bidirectional LSTM, Gated Recurrent Unit, Bidirectional Encoder Representations from Transformers, and ALBERT. It is observed that by composing the word embedding with the neural network models there is no need to construct any particular attribute or feature. By using word embedding, the bio contents are formatted as dense vectors which are fed as input into the neural network models as a sequence of vectors. The scores shows that our proposed approach has outperformed the traditional machine learning techniques for detecting medical occupations among Twitter users. ALBERT performs the best among the deep learning networks with an F1 score of 0.90. Overall, this chapter presents a novel method for detecting the occupations of Twitter users engaged in the medical domain by merging word embedding with state-of-art neural networks. The outcomes of our approach demonstrates that our method can further advance the process of analyzing corpora of social media without going through the trouble of developing computationally expensive features. Immediately following Chapter 4, this thesis will present an extension of this work that is currently submitted in a journal.

Chapter 4 presents a link prediction model trained on the follower and following IDs of Twitter users identified in the previous Chapter. The model is based on the homophily concept, that nodes that are highly connected in a network graph belong to similar communities or clusters that are embedding very closely to each other. Link prediction is mainly a

task of computing the probability of whether an edge exists in a particular network. Traditional methods calculate the similarity between two given nodes in a static network. Recent research has focused on evaluating networks that evolve dynamically. Although predicting links in dynamic networks is a challenging task, deep learning techniques and network representation learning algorithm, such as Node2vec, have shown remarkable improvements in predicting links. However, the Stochastic Gradient Descent (SGD) method of Node2vec is vulnerable to fall into local optimum and as a result Node2vec fails to capture the global structure of the network. To tackle this problem, we propose integration of NOde2vec anD Deep Learning mEthod (NODDLE) (integration of NOde2vec anD Deep Learning mEthod), a deep learning framework in which we merge the features extracted by Node2vec and feeding them into a four layer hidden neural network. NODDLE takes advantage of adaptive learning optimizers such as Adam, Adamax, Adadelta and Adagrad for improving the performance of link prediction. Experimental results show that our method yields better results than the traditional methods on various social network datasets.

Chapter 2

Background

All of this chapter has been submitted as the following peer-reviewed journal article:

- Khanam, K.Z., Srivastava, G., & Mago, V. (2020). The Homophily Principle in Social Network Analysis: A Survey.

Over the course of my degree, I studied and surveyed the effects of homophily in social networks, as well as summarized the state-of-the-art methods for identifying and measuring certain effects in a variety of social networks that have been proposed in recent years. Finally, this study has been wrapped up with a critical discussion of open issues and future research directions.

2.1 Introduction

Homophily is a well-established phenomenon that has been observed to occur frequently in social networks, where users with similar contexts have a nature of connecting with one another constantly, and this principle is also a meticulously thought-out field in the domain of social sciences [117, 141, 209, 68, 70]. Homophily is a social concept where people's personal networks tend to be more homogeneous than heterogeneous such that the communication between similar people occurs more frequently than with dissimilar people [122]. The main driving forces for initiating these networks are social influence and homophily. In other words, the importance of establishing connections between people does not rely upon 'what you know' but 'who you know.' In order to study this phenomenon, various studies have been conducted by sociologists on multiple socio-demographic dimensions of race, age, social class, culture, and ethnicity. For example, friends, colleagues, spouses, and other associates are inclined to mixing with each other who are similar to them than with randomly selected members of the same population [122, 192].

Studies in homophily usually have been conducted by surveying a group of human subjects which in most cases belonged in a specific geographical location [31, 189, 6, 55]. For example, one study showed that American high school students have a tendency to make friends with other students that belong to the same race and gender [130]. Initially, homophily was classified into two categories - status homophily and value homophily [103, 96, 183]. Status homophily mainly focused on the social position of the individuals inferring that individuals belonging to similar social conditions are inclined to mixing with one another. Value homophily in contrast is based upon the similarity of thoughts of individuals leading to the belief that individuals with homogeneous thoughts are inclined to connect with others even though differences may lie in their social positions [47, 134, 51]. Although, researchers

have successfully conducted experiments with human beings, the results were often based upon real-world scenarios of only small groups [177]. In order to fill the gap in the analysis, social media platforms come in handy as social networking sites such as Twitter and Facebook have become extremely widespread, with over 126 million daily Twitter users [148, 86] and Facebook having approximately 1.2 billion daily users [56, 180]. Reactive interfaces like those available through social networks provide users with the opportunity to be more open about their opinions, perspectives, thoughts, likes and dislikes [111, 90]. As a result, social media platforms are becoming more and more popular among users [15, 132]. These platforms are known to help users feel more involved. Users feel that they are able to participate in things that are happening around the world. Furthermore, such platforms help users in raising their voice against unjust acts or issues [59]. Therefore, both status and value homophily have been analyzed recently in social networks in order to evaluate whether these types of homophily phenomenon exists in these types of networks. Moreover, if homophily exists, whether it increases or decreases in digital environments has been studied [135, 186, 16].

The effect of homophily has been vastly studied in different types of social media data. From textual data (Twitter tweets) to follower lists of online social accounts [141]. However, no detailed survey has been conducted to date based on the works of social media networks related to the homophily principle. Therefore, the main aim of this chapter is to focus on providing a thorough review of the related works conducted on social media networks based on the homophily principle.

The rest of the chapter is organized as follows. Section 2.2 presents the methodology that has been used to extract high quality articles in order to conduct the survey. Section 2.3 discusses the role of homophily of the various ways in which the homophily effect has been analysed in multiple domains of social media data. Section 2.4 discusses on the predictions

made in many fields of social network by using the homophily effect. Section 2.5 introduces a comparative study of the social network analysis, conducted by measuring homophily in multiple applications, the different types of network models constructed in each of the proposed models. Section 2.6 includes the different types of datasets used to validate these proposed, homophilous models. Section 2.7 discusses about the state-of-art methods used for detecting homophily in social networks, the limitations of these approaches and directions for future research. Fig. 2.1, shows the overall structure of the chapter.

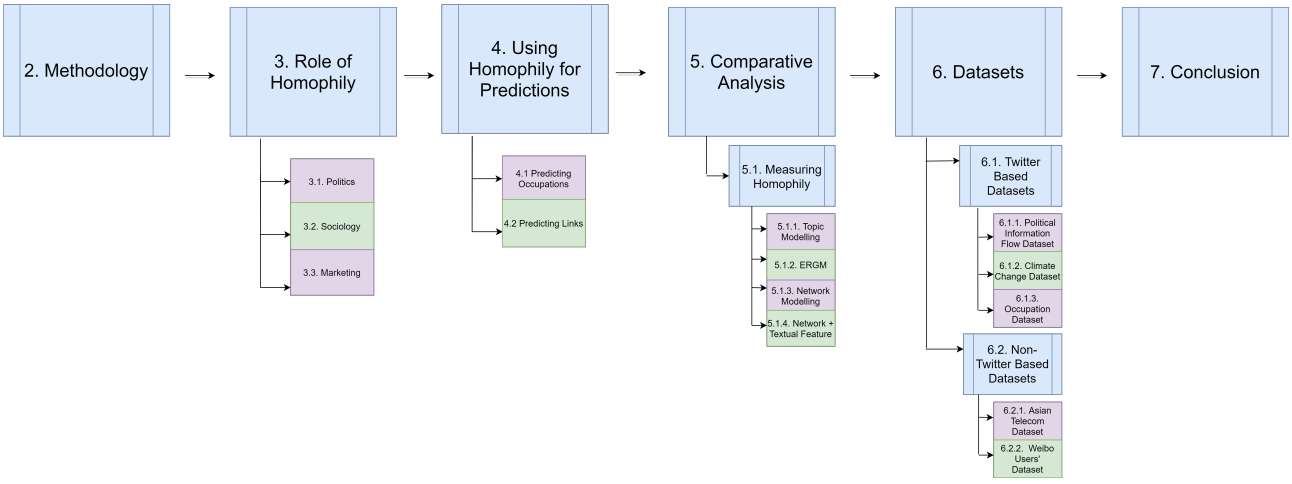


Figure 2.1: Overall structure of the background chapter

2.2 Methodology

Keywords, such as Homophily, social media, and degree distribution have been used to search for papers related to analyzing homophilous models. However, it is not only important to find the appropriate papers based on keywords but also to extract papers from top venues. As such articles have a high impact factor. As a result, the h-index of the venue, where the paper was published and the number of citations of the paper were considered. We have

mainly focused on the venues which have an h-index of 50 or above, from Q1 or Q2 journals, and the articles having a minimum of 100 citations. Using this information[1], papers for this survey were obtained by accessing them through university library resources. Recent surveys [113, 84, 63] have also reported adopting the similar approach. Fig. 2.2, shows the h-index of the articles cited in the survey, we can see most of the articles' h-index is from 50-100. As h-index is a venue-level metric which is used to evaluate the impact factor and citations of the publications of the venue. Thus, Fig. 2.2 shows that most of the articles selected have high h-index. The table A of Appendix A shows the articles selected for this survey with the venue, number of citations, quartile, h-index, as well as the year of publication.

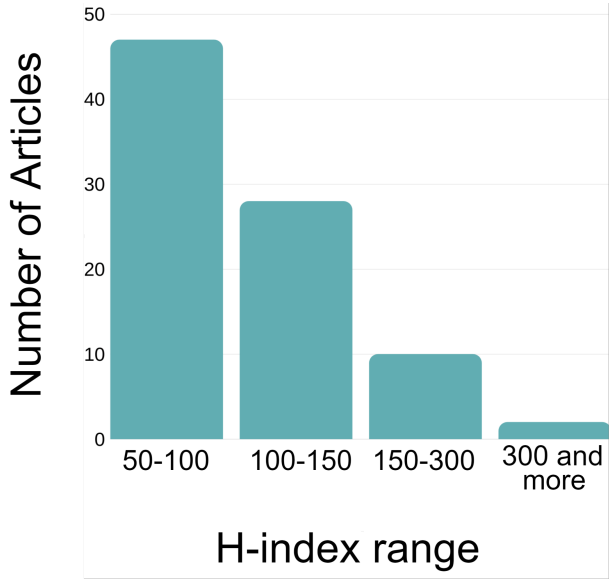


Figure 2.2: The number of articles' H-index from a range of 50 and above

An article from a high h-index venue, with a high number of citations shows that the paper is reliable and trust-worthy for the academic community. Fig. 2.3 shows the total percentage of journals, conferences, and other types of articles such as book chapters, workshop papers

[1]https:/www.scimagojr.com

that have been cited in this chapter. It can be seen from Fig. 2.3 that most of the articles selected for this survey are from journals. Furthermore, in Fig. 2.4 we can see that majority of the articles are taken from Q1 journals. However, for some of the articles, the information about belonging to certain quartile was missing. For such cases, we have only focused on the remaining metrics. Besides, the papers selected were from 2015 onwards so that the approaches used in the recent papers could be studied more exhaustively [108, 200]. However, if any articles have major contributions, such as introducing novel algorithms or approaches used in measuring the degree of homophily, then, they are considered for this survey. This is because homophily is not a recent concept and the impact of these papers is more important than the year of publication. Fig. 2.5 shows that most of the articles have been selected from 2015 onwards.

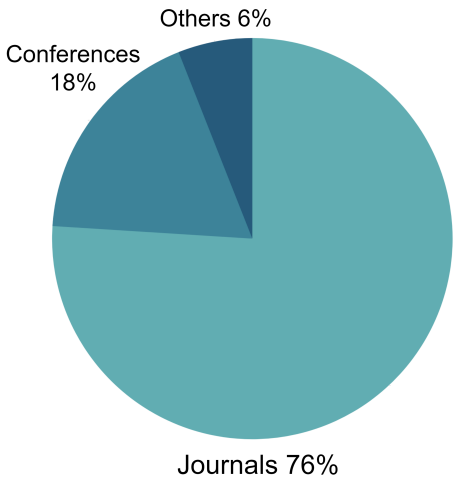


Figure 2.3: Percentages of Journals, Conferences and other type of articles cited in the survey chapter.

A word cloud, as shown in Fig. 2.6, was generated from the abstracts of the papers selected to get a visualization of the most important word in the field of homophily [83]. We

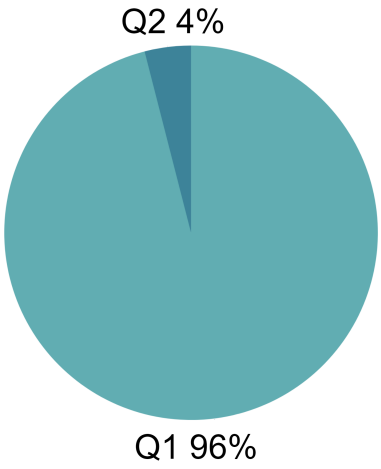


Figure 2.4: Proportion of Q1 and Q2 Journals referenced in the survey article.

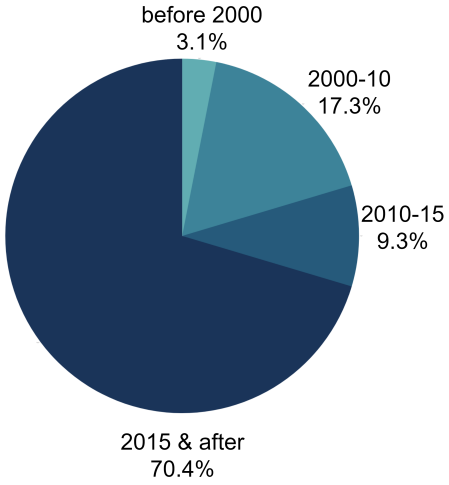


Figure 2.5: Percentages of the Year of Publication of each of the articles.

implemented a simple python code to form the word cloud. The abstracts are pre-processed by converting the text to lower case, removing the punctuation, and commonly used English stop words, available in the nltk library. Then the word-cloud is built using a word cloud library. The importance of each word is represented with the font size and color. The

darker the color of the word, the more significant the word is. The larger the font size of the word, the more frequently the word has occurred in the abstracts. Fig. 2.6 shows that the homophily word has the largest font which depicts that most of the papers were about homophily related issues. Moreover, the word Twitter and social network has the second-largest font [159, 32]. The color of the font is darker which shows most of the papers discussed about social media. Moreover, the Fig shows that words like social, graph, networks, echo, and homophily are highly co-related to one another. The word *echo* is used, in social media, when users raise their voices and talk about politics or debating about a particular issue. As a result, it can be concluded that most of the abstracts talk about visualizing social network graphs from the issues discussed in social media platforms by using the homophily concept.

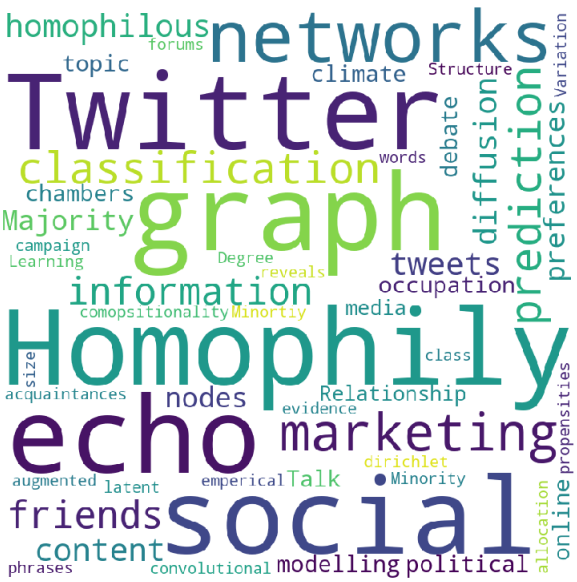


Figure 2.6: Visual representation of the important words from the abstracts

2.3 Role of Homophily in Social Media

The Internet has the ability of connecting people, with all kinds of interest, all around the world. As a result, it can be assumed that when social ties are formed between individuals in social media, homophily will be less likely to appear among users. However, homophily has rather shown to increase in social media platforms [197, 25, 97, 198]. For instance, earliest research has highlighted that, when people used Microsoft Instant Messaging, users were more interested to converse with others belonging to a equivalent ages,location and native language [106]. Moreover, the more individuals communicated with one another, the more related their online searches were [167]. Similarly, increased homophily was also detected among groups of Facebook friends, where they had similar thoughts on ideology or political orientation [121]. The effect of homophily has been studied in various domains to observe if homophily has any effect on social networks or not. For example, Twitter is a popular micro-blogging platform that has been considered as an effective tool for studying the interactions between the social media users [98, 123]. Homophilic studies have been conducted at an exhaustive rate in the domain of politics, marketing, and sociology as well. Insightful information has been extracted from these approaches. The following subsections discuss in detail the approaches that have been proposed in these domains.

Politics

In the field of politics, homophilic studies range from analyzing the users participating in political debates to observing the network of politically engaged users on a variety of political activities [68, 186, 72, 193]. On a politically oriented website named ”Essembly”, users were observed to form positive and negative ties with people having parallel thoughts and different reasoning on an ideology respectively [74]. Furthermore, when smaller networks

were studied more in-depth, several characteristics such as gender, age and level of education have proven to be strong predictors with network structural characteristics. Moreover, these characteristics were used for investigating the existence and strength of positive ties among individuals [60].

Homophily principle was also used to study the flow of political information among the majority and minority groups on the Twitter platform [68, 72, 99]. Recent studies have shown that the majority of larger groups received political information more quickly than smaller groups. Both groups were exposed to similar political information and it has been observed that the flow of information was faster among the larger group. Substantial evidence of homophily was detected, when users following a specific political party were more likely to connect with other users following the same party. The flow of information through the social media network was faster among the majority groups since they had more network connections and so they received more information at a faster rate. To sum up, political information is considered as an extremely influential information. Therefore, increasing exposure to such information among the like-minded majority (large) users can further increase political divergence among the users.

Furthermore, social networks of users exchanging views about global warming on Twitter were examined. The users' attitudes towards global warming were classified based on their message content [186, 82]. The social networks were categorized by opinion-based homophily and the users were manually labeled as "skeptic" and "activist" groups based on their message content. Results have shown that, users generally communicate only with other similar-minded users, in communities that are influenced by a common view. Moreover, the messages of like-minded users have shown to be a positive sentiment in most of the cases, whereas, messages from skeptics and activists held a more negative comment. Overall, discussions of climate change in social media often take place in the polarising "echo chambers" where

political issues are discussed, and also in “open forums” and mixed-opinion communities [143, 60].

Sociology

Homophily, meaning ”love of sameness”, is considered to be a sociological theory that like-minded people will be inclined towards each other and will have a tendency to act in a similar way [103]. This behavior of individuals has been studied in social media platforms as well. Social media generally consists of majority and minority groups, where the majority group is considered to have stronger connections with one another in its group and also tends to have higher network communications [68, 16]. Compared to the majority group, the minority group not only has fewer members in its group but is also deprived from receiving information quickly [85]. Thus, the relationship between majority and minority group in social media is studied in depth to observe how the groups and the size of the groups are formed and the groups react to one another in social network [85, 71].

In order to study the influence of homophily between the minority groups, the levels of homophily was calculated by combining the centrality measures with the preferential attachment network [12, 146, 85]. The model focused on multiple ranges of homophily and density of populations by capturing the degree distributions and ranked the minority groups in empirical social networks of scientific collaboration and dating contacts[85]. Experimental results have shown that as the volume of the minority group decreased, the heterophilic interactions were greater than the homophilic interactions. However, multiple assumptions were made. For instance, all the members of the minority groups were considered to be equally active and behave in a similar pattern and the group size differences were omitted. These factors can cause a bias estimate in the ranking of the groups. A major drawback

was also faced when validating the proposed model such as finding adequate numbers of large scale data representing the minority groups, since, remote and hard-to-reach minority groups are often absent from the social network datasets [164].

Marketing

Comparative analysis has also been conducted on homophily and social influence effects on product purchasing [117]. This analysis examined problems related to whether a company should target customers based on homophily or the social influence effect. If a company relies on the homophily principle then they target the existing customer's friends directly as they have a tendency to purchase similar products. Whereas, if the firm emphasizes on social influence of the existing customers then they only target the existing customers and rely on them to promote to their social circles. Therefore, for cost effective marketing strategies, it is extremely important to separate these two effects. However, such approaches are challenging since both phenomenons end up producing similar outcomes. As a result, a product choice model has been designed via the hierarchical Bayesian model which was implemented with a dataset that consists of both communication and product purchase information over a three month period provided by Asian Telecom Company [117]. A strong homophily effect was detected on the choice of products. When one of the factors was ignored in the Bayesian model, it resulted in an overestimation of the other factor and this shows that social influence and homophily effects are highly connected with each other. Ignoring any of the factors leads to biased estimates in the Bayesian model. Furthermore, as network structures are versatile in nature [54], it was difficult for the model to detect strong and weak ties in network structures. This is because, some people in the network might have many friends with weak ties to one another consequently others might have few friends but with extremely strong

ties. As a result, the model can be further improved to inspect the strength and the impact of social ties with respect to a customer's decision on product purchasing. This will help to identify customer preferences in such versatile networks. Thus, an improved model is required that can further differentiate the effects of homophily and social influence.

Gender homophily was also invested at a global scale in online book market such as the amazon.com in order to address the research question that whether readers are more biased to reading books from the same authors or did diversity exist in book selections. In online book markets, book sales lead to important ties among the books. These book ties create large book networks that model the collective information about the reading habits of the customer. The study was conducted with large volume of dataset that had records of the books sold to readers to investigate gender homophily on a large scale. Assortative measures such as Newman's degree based Assortative co-efficient metric was used to measure homophily between gender-based readers and authors. Assortativity or Assortative Mixing is defined as the tendency of a network's node to connect to other nodes that behave in similar patterns. Their findings have shown strong homophily in the book networks. Particularly, readers that prefer reading female authored books tend to buy other female books. While the male authored book readers tend to buy less female authored book compare to male authored books. Thus these findings have revealed the essence of gender homophily. However, their study had some limitations such as finding the name of the main author as many books has non-english author names and it is difficult to figure out the gender from non-english author names.

Research was also conducted to study the presence of homophilic patterns based on the usage of hashtags in a Twitter mention network based on a Cause-Related Marketing (CRM) campaign [188]. CRM is a mutually beneficial collaboration between a corporation and a nonprofit organization. It is designed to promote social responsibility in the public

community. However, CRM is a risky and controversial issue since this campaigning varies from receiving skepticism to full support of the customers [13, 133]. Gillette's CRM campaign "The Best Men Can Be" was used, to test the hypothesis that whether homophily exists in such marketing campaigns or not [188]. The brand's goal was to address concerns based on gender inequality and bullying of men and encourage a better lifestyle for youngsters. The company, moreover, guaranteed to make a donation of 1 million to NGOs fighting for gender inequality. When the campaign ad was released, the ad received positive feedback from some of the customers because of its positive message [89]. However, others felt that the ad was a bit offensive representing men as sexually harassing, and bullying. Thus, it received negative feedback from the rest of the customers [21]. Hence, two groups were formed in the social media where one group was supporting the cause while the other group opposed its motive. As the brand is well-known and discussions on this campaign became a trending topic on Twitter. Thus, this CRM campaign was an ideal fit to analyze how the users communicated and reacted with other users on the Gillette's ad. For CRM's marketing campaign, topic modeling was used for extracting information [142]. Topic modeling was conducted on 100,000 original tweets, profiling the topics related to the CRM campaign tweets [22]. Based on the users' engagement, the network of the CRM event's related hashtags were analyzed with the aid of Exponential Random Graph Models (ERGMs) [168]. ERGMs are statistical models that are commonly used for analyzing data regarding online social networks [172, 44]. Results generated from this model showed an increased tendency of homophily on the network of users. The degree of homophily inspected was based upon the common views of the users. The results of topic modeling on Twitter have revealed that users are highly dependent on established social networking platforms to discuss important issues [89, 64]. Furthermore, users tend to react more to the tweets of influential, popular users which enable the users to be more reactive during online discussions. Moreover, these users

showed homophily on the usage of hashtags. Thus, ideological hashtags served as measures of homophily as ideological hashtags refers to a person's identity and thoughts [23]. One such example of ideological hashtag is the usage of *#BlackLivesMatter* or *#AllLivesMatter* which reflects the user's ideological position is based on the social justice issues to a large extent. Therefore, hashtags not only express a user's self-identity but also helps similar users to identify and connect in a versatile community [21, 9].

2.4 Using Homophily for Predictions

Homophily concepts can be implemented for predicting certain features. For example, homophily principles have been applied in the link prediction area for studying the probability of one user to be connected with another user. On the other hand, homophily in Twitter has also been examined to predict occupation of users based on the information of the users' followers and followings' IDs. The usage of homophily concept in predicting certain features are discussed below.

Predicting occupations

In order to predict the occupation of Twitter users, the homophily principle was used to conduct social network analysis on the biographical content of the user's follower/following community [141]. Occupation prediction is considered as a multi-classification problem since the model is specialized in predicting multiple occupational classes. Furthermore, the results concluded that a user's follower/following community provides insightful information for identifying the occupational group of each of the users. The model was designed with Graph Convolutional Network (GCN) [95] which has enhanced the model to work efficiently by training on only a small fraction of data. Graph convolutional network is a recently proposed

graph based neural network learning model, which specializes on learning graph-structured network data [95]. Thus, by using the homophily principle and GCN, a better result was achieved for predicting occupation class with an accuracy of 61%. A similar work was done to predict the occupational class of Twitter users where the dataset contained the historical tweets of the users [149], however, an accuracy of only 50% was achieved.

Predicting Links

The homophily principle was used as motivation in various works of link prediction, where link prediction is calculated based on the similarity between two entities.As a result, it can be used to predict future possible links in social networking platforms [46]. For example, researchers have proposed a model to investigate the associated links of document's topic distribution between people discussing about related topics. This study shows how topic distribution is mainly affected by the distribution of the topics of its nearest neighbors [190, 93, 188].

Particularly, a joint model was proposed in which link structure has been applied to define clusters. Here, each of the clusters was allocated with its own segregated Dirichlet prior for topic distribution. Using such priors have shown to be very helpful as in previous works only document priors were applied [207, 128]. Discriminative and max-margin approaches [207, 208] have been used for designing the contextual documents and generating good link predictions. Moreover, lexical terms have been used in the decision function in order to improve the strength of the prediction [137].

In summary, users in social media are not only comfortable at expressing their self identities but also have a tendency to connect with one another, with similar interests, in a versatile community. Several studies have been carried out to study the homophilic patterns

especially among the majority and minority groups. Studies have deduced that the majority and larger groups receive information faster than smaller, minority groups and information reaches like-minded individuals more quickly [119, 80, 68]. Whereas, the minority groups are deprived from receiving information instantaneously as such groups have fewer members in their groups hence have fewer network connections [85]. On the other hand, a strong homophily effect was detected on customers having similar product tastes and based on the attraction of users having a common view [117, 186, 188].

2.5 Comparative Study of Related Works for homophily detection

In this section, comparative analysis has been performed on various approaches proposed in order to detect and calculate the level of homophily. In addition, the network and language models defined by each of the state of art methods are also explained in details.

Measuring the degree of Homophily

Multiple approaches have been proposed for measuring the level of homophily with the aid of topic modeling and network modeling. Recently, the degree of homophily has also been calculated by combining both textual as well as network features by using a highly efficient neural network model that has outperformed the existing traditional methods [141, 85, 68, 188].

Topic modeling

Topic modeling is considered as an unsupervised machine learning technique which does not need the aid of humans to determine the topic of a set of documents [18]. It can automatically detect the main theme of given paragraphs or documents by clustering groups of words or similar expressions that best portrays the set of documents or paragraphs. The topic model proposed based on the homophily concept specializes in detecting high-quality topics to test the hypothesis that whether people talking about similar topics are connected with each other or not [190]. For example, the Latent Dirichlet Model (LDA), is a topic model, that maps the documents to the topics based on the distribution of words [22]. LDA model was modified with the concepts of homophily to not only detect the high-quality topics but also predict whether the people having similar posts in social media are connected with one another or not [22].

Generally, the most frequent words of each of the documents are aggregated into K words clusters by using the k means algorithm [118]. Thus, for any word token $w_{d,n}$, for the word token belonging to a cluster k, any other token $z_{d,n}$ being a neighbor of $w_{d,n}$ will also belong to cluster k. The d represents the document and n represents the number of documents. Therefore, in order to find the topic k's major words, skip-gram transition probability [125] is calculated for each $w_{k,i}$ word as in Equation 2.1.

$$S_{k,i} = \sum_{j=1, j\neq i}^{N_k} p(w_{k,j}|w_{k,i}) \tag{2.1}$$

where, N_k indicates the number of words in topic k, words with the highest probabilities are used as the designated topic words for each of the documents in the sample. Regression is used to compute the topic distribution between d and d', for predicting the link between the two documents, which is depended on the similarity of their topic patterns. Therefore,

the regression value is defined in Equation 2.2.

$$R_{d,d'} = \eta^T(\overline{z_d} \circ \overline{z_{d'}}) + \tau^T(\overline{w_d} \circ \overline{w_{d'}}) \tag{2.2}$$

where, $\overline{z_d} = \frac{1}{N_d}\sum_n z_{d,n}$. Similarly, $\overline{w_d} = \frac{1}{N_d}\sum_n w_{d,n}$; $\circ$ denotes the Hadamard product [75]; η and τ are the assigned topic based weight vectors and document link predictions.

In some studies, the perplexity metric was used as a measurement for evaluating the model's topic modeling performance [57, 24, 45]. Perplexity is a measurement based on the quality of a probability model predicting a sample. Results have assured that the proposed model outperformed the traditional LDA model for topic modeling. Furthermore, for validating the model in terms of its performance for predicting document link prediction, the Predictive Link Rank (PLR) metric was used. PLR outputs the average rank of a document with the documents to which it has been linked with. High training performance was achieved that showed user interactions can contribute to better link prediction. However, the testing performance score was much lower than the training performance score. This shows that the model has over-fitted since the model could not perform well with the testing dataset. Even though the new model outperformed the traditional LDA method, for document link predicting task the overfitting issue was not resolved.

LDA-based topic modeling focuses on topics co-occurring frequently. However, the main drawbacks of LDA based approach is the need of specifying the "appropriate" number of topics that the LDA has to predict [22]. Statistical indices have been proposed to address this issue [10, 30, 48] which includes differentiating each pair of the topics by the difference of each pair of topics or their distance. Although, these methods can approximately calculate the number of topics in a given corpus, proper gold standards or benchmark still does not exist. Therefore, human interpretation is still required for rendering the topics into

the unsupervised topic modeling method [34]. Most importantly, the performance of these methods are not analyzed in documents such as a tweet which consists of only sentences with few number of words. Thus, for measuring the homophily of users using similar hashtags in Twitter posts, the co-occurrence of the topic was calculated by using Mimno *et al.*'s approach [188, 129]. Mimno *et al.* state that for any document, the leading words in a topic profile are likely in the same document. The coherence in Mimno *et al.*'s approach is defined as:

$$C(t; v^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{2.3}$$

In Equation 2.3, $D(v)$ denotes the frequency the document, word v and $V^{(t)}$, topic t has a list of M words.The topic model can predict better when the result has an output that is close to zero. [150]. In order to evaluate the quality of the predicted topics, the distance between each of the topic was calculated using Jensen-Shannon divergence [150, 171].

Exponential random graph modeling

Exponential random graph models were used to evaluate whether the networks of users engaged in conversations influence the users' response in online discussions [188]. In ERGMs, nodes' connections at the same degree is considered to be an indicator of users conversing frequently [131]. These graph models are statistical models that are composed of network structures [116]. ERGMs have been used in multiple domains of the social networks such as social media settings and to study communication between the users. For modeling the presence of any ties that may exist between the network's local and structural factors [62, 158, 162].

Generally, the ERGM model is designed by aggregated tweets and hashtags which are posted by the users. [188]. The top hashtags are categorized as either conceptual or ideo-

logical markers based on the definitions provided by Blevins *et al.* [23]. Ideological hashtag refers to the identity or identification, perspectives. On the other hand, conceptual markers are considered as personal thoughts on particular events. Then, the users mentioned in the network as @user and the classified hashtags are fed into the statistical modeling as nodal attributes. The mention network is modeled because it helps to visualize a discussion that starts by actively exchanging information with one another rather than relying on what others have said.

Network modeling

Network modeling is a flexible way of representing a group of connected objects. The objects are represented as nodes or vertices and the connection between the nodes are represented by edges. Network modeling is generally used to visualize the various types of networks. These models are constructed from the social media data consisting of users and how the users are connected. Many researchers have claimed that the flow of information in social media is dependent highly on the majority groups where more users are connected. Moreover, the majority groups have a larger social circle compared to the minority groups. The reason behind having a larger social circle is due to homophily [68, 85].

The homophily among majority groups was studied between a network of politically engaged users of Twitter [68]. Due to the shortage of measuring the political orientation and ideology of the Twitter users directly. The research emphasized on the users following politicians from the two major parties- *Conservatives* and *Liberals* of the House of Representatives, in the 2012 general election. In order to analyze the degree of homophily, the individuals are divided into two groups - conservatives *(C)* and liberals *(L)*, depending on which political party each group supports such that: $t \in C, L$. The group sizes are normalized and symbolized as wt which is the weight of the tweet, where, $w_C + w_L = 1$. Conservatives was

randomly selected as the majority group and liberals was considered as the minority group. Therefore, $w_C \geq 0.5$.

When two individuals belonging to the same group are randomly selected, then the probability of the two individuals communicating with each other is denoted by π_s and the probability of the two individuals connecting with each other belonging to different groups is represented by π_d. Moreover, it is also logical to assume that individuals belonging to the same group have a higher tendency to interact with each other compared to two individuals of two different groups communicating with one another. As a result, $\pi_s > \pi_d$. Thus, an individual belonging to a group t will be having similar $\pi_s w_t$ interactions, and $\pi_d(1 - w_t)$ different interactions. Therefore, in this study, the Homophily principle has been evaluated as such:

$$H_t = \frac{\pi_s w_t}{\pi_s w_t + \pi_d(1 - w_t)} \tag{2.4}$$

In Equation 2.4, the greater the value of $\pi_s w_t$, the higher the degree of homophily. As a result, if conservatives are more prominent and links are formed with similar types, then conservatives would tend to be more homophilous in nature. Similarly, the liberals would be more heterophilous. Furthermore, the time taken to reach the information to majority and minority groups were also taken into account. So it was considered that each user produces information with probability of ε at time $\tau = 0$. Bass model or Bass Diffusion Model was used in order to generate the proposed model [17]. The Bass model uses differential equation that describes the procedure of new products getting adopted in a population. In this case, the product is the political tweets. If interaction occurs between two users, then the user exposed to information transfers the information to unexposed user with a probability of q. Hence, by following the Bass model, the rate of information diffusion is defined as such:

$$F_t^\tau = F_t^{\tau-1} + (1 - F_t^{\tau-1}) f_t^\tau \tag{2.5}$$

In Equation 2.5, F_t^τ is defined as the fraction of group t receiving information at time τ which is then connected to the fraction transmitted information at time $\tau - 1$. F_t^τ is the chances of group t getting information at time τ if not being exposed to information at time $\tau - 1$. Therefore, f_t^τ is defined as:

$$f_t^\tau = qw_t\pi_s F_t^{\tau-1} + q(1-w_t)\pi_d F_{-t}^{\tau-1} - q^2 w_t(1-w_t)\pi_s\pi_d F_t^{\tau-1}F_{-t}^{\tau-1} \tag{2.6}$$

The symbol $-t$ in Equation 2.6 refers to the other group. The first term denotes the likelihood of receiving the information from the individual belonging to the same group. Second term denotes the likelihood of receiving the information from a different group. Third term refers to the likelihood of both groups receiving information. In conclusion, if biased interactions are present, $(\pi_s > \pi_d)$, the majority group member will receive information faster than the minority group $(F_C^\tau > F_L^\tau)$ for every τ times. Based on receiving a higher probability score for the majority group, it is deduced that homophily is directly proportional to the rate of flow of information among the users. Moreover, the diffusion of information is relatively uniform with groups having a higher number of connections based on having similar political orientations. Thus, larger groups are exposed to information at a faster rate. Therefore, a close relation of homophily and diffusion of information is shown in this approach.

A similar approach was also used to measure the level of homophily between the minority groups [85] which is shown in Equation 2.4. The homophily was used as an additional parameter in the famous model of preferential attachment proposed by Barabási and Albert [12]. Preferential attachment means that a node is more likely to receive new links if it has higher number of connections. Thus, such nodes are more powerful since they can tightly hold links with one another. The growth of complex evolving networks was calculated using

the fitness model which is based on the Barabási–Albert model [85]. In this model, nodes with different type of characteristics can grasp links at different rates. Hence, the fitness is calculated by the degree distribution of each of the nodes. This is how the model can predict a node's growth. In the Preferential Attachment model, at each step, a new node which just approached, its degree and group attachment is calculated for the possibility of the node to be attached to the pre-existing nodes. The chances of node j to be connected to node i is defined as:

$$\Pi_i = \frac{h_{ij}k_i}{\sum_l h_{lj}k_l} \tag{2.7}$$

In Equation 2.7, k_i generally, represents the degree of node i and h_{ij} is the similarity between nodes i and j. The similarity between each of the nodes is build, based on the nodes' attachment when the network was generated. If by any chance, the new node is not confronting individuals from the same network, it can stay deserted till the node confronts a newly approached node that is coming from the same network.

On the other hand, when the online debate on climate change was studied, the degree of homophily among the individuals was measured on the number of times the edges were connecting users on homogeneous/heterogeneous views[186]. The high frequency of edges between the homogeneous users, and similarly, the low frequency of edges between the heterogeneous users were considered as the measure of homophily. The probability of picking node i as the root or focus node for a given edge, were denoted by:

$$P_{source}(i) = \frac{k_{out}(i)}{\sum_{j \in a,s} k_{out}(j)} \tag{2.8}$$

$$P_{target}(i) = \frac{k_{in}(i)}{\sum_{j \in a,s} k_{in}(j)} \tag{2.9}$$

In Equations 2.8 and 2.9, k_{out} is node out-degree and k_{in} is node in-degree. This mathematical technique generates nodes' networks with homogeneous degree distributions.

Other type of metrics that are used to measure the quantity of homophily includes the Newman's global assortativity coefficient [136]. Assortativity or Assortative Mixing is defined as the propensity of a network's node to connect to other nodes that behave in similar patterns. Newman's global assortativity coefficient is a vertex degree metric used to identify whether vertices in the graph tend to mix with homogeneous or heterogeneous vertices. The global assortativity coefficient is measured interms of the discrete characteristics for each of the vertex. The coefficient gives values within the range [-1,1]. The coefficient is zero for a given network if the vertices connect randomly. A high positive value indicates that the high degree vertices links preferentially with other high degree vertices and the other way around for the low degree vertices. The assortative metric was used to measure the degree of gender homophily in online book networks [27]. This coefficient is mainly applied for evaluating that whether customers from the same gender have the tendency to buy books written by the same gender author or are the books bought randomly by chance.

$$r^k = \frac{\sum_i e_{ii} - \sum_i a_i^2}{1 - \sum_i a_i^2} \tag{2.10}$$

Equation 2.10 shows how the Newman's degree based assortivity has been defined for calculating homophily for the given situation . For a given book graph G^k, the assortativity coefficient r is denoted by i, j, where $i,\ j \in \{$*male first author, female first author, book collections*$\}$. Out of all the edges in G^k the portion of edges that link two book from any of the gender categories i and j is depicted as e_{ij}. The fraction of occurrences same gender edges in G^k is denoted by the term $\sum_i e_{ii}$. The a_i term represents the portion of edges in which one of the end is a book from category i gender. For a randomly connected graph

$\sum_i a_i^2$ term shows the portion of same gender occurring edges. The normalized differences between the between the real ans random fraction of same occurring gender edges in graph G^k is denoted by r^k.

Combining Network and Textual features

Recently, a new neural network model has been proposed known as the Graph Convolutional network (GCN) [95]. GCN has been used for extracting the textual and the network features for identifying the homophily connection between the Twitter users and their followers/followings list [141]. GCN has not only enabled the model to achieve a high performance, but also the model was successfully trained with only a fraction of data. GCN graph-based neural network model *f(X,A)* with layer-wise propagation rules is defined as such:

$$\widehat{A} = D^{-1/2}(A + \lambda I)D^{-1/2} \tag{2.11}$$

$$X^{l+1} = \sigma(\widehat{A}X^l W^l + b^l) \tag{2.12}$$

In Equations 2.11 and 2.12, X denotes the matrix of the features for each of the nodes(users). X^0 is the initial feature with a input size of *(nodes * features)* and A is the adjacency matrix of the dimensions *(nodes * nodes)*. D represents the degree matrix of $A + \lambda I$, where λ is the hyperparameter that controls the weight of the node among its neighbor hood. W^l , b^l are the trainable weights and bias for the l^th layer. In each GCN layer, nodes accumulates its closest neighbors' features by linearly converting the representation using weight *(W)* and bias b respectively. σ represents the activation function used in the GCN model for optimizing the performance. Then, the number of GCN layers determines the path of the node from its closest neighbors' features.

The inputs of the adjacency matrix $\widehat{A}$ are all the network IDs (target users and their followers and following list IDs). A feature matrix of the biographical descriptions of the each of the target users' followers/following lists. This is because the biographical description of each of the target users' followers/following lists might be similar with the target users. Pan et el. claimed that an accuracy of 61% was obtained, which outperformed the results of the existing methods [141]. Thus, GCN was able to extract the rich network and textual information in order to learn the homophily connection between the users. However, the model was trained with only a small fraction of data and thus if a larger amount of data would be used the model would achieve a better result for predicting the occupation of target users.

In summary, various types of models have been proposed for measuring the degree of homophily. Mainly network modeling and topic modeling have been used to find out the strength of a relationship of a user with the user's nearest neighbors. To also investigate, how many users are involved with one another. Network modeling focused on the network features which involved calculating the number of connections each user has and the strength of the connection of the users with one another. On the other hand, topic modeling focused on clustering the main topics of each of the users, based on their textual tweets and how similar the usage of topics are with one another. Recently, the degree of homophily has also been evaluated by incorporating textual and network features with highly efficient neural network model that has outperformed the existing traditional methods for multi-classification problems.

2.6 Datasets

Multiple datasets have been used for measuring the degree of homophily in multiple fields such as link prediction and flow of information among majority and minority groups [141, 68]. For example, the dataset provided by Asian Telecom Company was used to analyze if the homophily effect can influence product purchasing [117]. This section discusses the various datasets used to study the homophily effect.

Twitter-Based Datasets

Currently, there are no standard datasets that have been used to study the effect of homophily. Rather in most of the cases, datasets were generated with the aid of Twitter Search API [138]. Relevant words or hashtags have been used to find tweets and hashtags among the users who have posted such tweets. For example, to search for tweets about climate change, hashtags such as *#climatechange* and *#globalwarming* are used to find such specific tweets about climate change along with the user information [186]. Based on this search, an extensive network is generated from these users. Different approaches such as network modeling and topic modeling are used to measure the degree of homophily. In Table 2.1:1, 2, 4, 5, 6, 9 show examples of some of the datasets that were generated to study the effect of homophily by using this approach, as the mode of data collection was very much alike. Hence, the description of the following datasets gives an overview of how networks are generated by using the Twitter Search API.

Political Information Flow Dataset

The twitter dataset for analyzing the flow of political information among majority and minority groups was constructed by targeting the politically engaged users [68]. These users

	Dataset	Source	Size	Public Private	Year	Advantage	Additional Information
1	Occupation Dataset	Twitter	34,630 unique users and 586,303 edges	Public	2019	Achieved the highest accuracy at predicting the occupation of 5000 target users	Predicted the occupation of target users based on the biographical content of the target users' social circles [141].
2	Political Information Flow Dataset	Twitter	2.2 million Twitter users with 90 million network links	Private	2018	Time taken for information to flow among the majority of voters could be efficiently calculated	Information flows faster among users with higher number of connection compare to users with less number of connections [68].
3	Majority-Minority Dataset	Generate artificial undirected network	The undirected network consisted of 5000 nodes and averaged over 20 simulations	Private	2018	Calculates homophily by combining the centrality measures with the preferential attachment network	Captures the degree distributions and ranks of the majority and minority in empirical social networks [85].
4	Climate Change Dataset	Twitter	590,608 distinct tweets from 179,180 distinct users was used	Private	2015	Hashtag analysis was conducted using mean Sorensen similarity [170]	Sorensen similarity could successfully detect homophily as greater values showed a greater constancy among a major population of active users [186].
5	Emotions and Political Talk Dataset	Twitter	70 datasets were collected based on 10 controversial topics, each dataset has tweets of 1500 users	Private	2016	Identifying the emotions based on political conversation by using k-mean clustering	Their findings suggested that oppositional tone were associated more with negative emotion clusters, while supportive clusters overlapped more often with positive emotion clusters [72].
6	Debate Dataset	Twitter	Collection of 900,000 tweets	Private	2019	Multi-classification problem of detecting arguments between users by training the model with labels - against, favor or none	Achieved f1 score of 0.60 by using Linear SVM method for predicting any argument occurring between users [99].
7	Asian Telecom Dataset	Asian Telephone company	300 million phone call histories of the company's approx 3.7 million customers	Private	2015	Hierarchical Bayesian model was developed with the communication information	Strong homophily effect was detected on product purchasing [117].
8	Weibo Users' Dataset	Sina Weibo website	Posts of 2000 users	Private	2015	Homophily is calculated by predicting links between the users' posts	The model could obtain better link prediction scores between the users by calculating the rate of similarity of each of the tweets [190].
9	CRM Campaigning Dataset	Twitter	100,000 posts from 75,302 unique twitter users	Private	2020	performed topic modeling on original tweets by using exponential random graph models (ERGM)	Strong homophily was detected among users using certain hashtags [188].

Table 2.1: Details of some of the datasets used to validate the presence of Homophily in the digital environment

were following at-least one account of a candidate running for the 2012 US elections. As a result, over 2.2 million users' data were collected from which 90 million network links were approximately identified. Users following more accounts of Republican political candidates than accounts connected with Democrats candidates were categorized as conservatives. Similarly, the users following more Democratic accounts were categorized as liberals. To measure the level of communication among the groups of supporters, approximately 500,000 retweets of the candidates' tweets and tweets that mention candidates were also collected and analyzed. The flow of political information among the groups of Twitter users was measured based on whether or not the users received a candidate tweet or mention through these networks. Moreover, the rate of information flowing through the political network was taken into account by measuring the time taken for these retweets to diffuse across the networks.

Climate Change Dataset

Twitter API was used to collect tweets between January 2013 and May 2013 that consisted of the trending hashtags on global warming such as- *#globalwarming*, *#climatechange*, *#agw (an acronym for "anthropogenic global warming")*, *#climate*, and *#climaterealists* [186]. Moreover, followers of each of the users posting such tweets were also identified. Hence, $590,608$ distinct tweets from 179,180 distinct users were used to generate the dataset. Hashtags were mainly used to search tweets as Twitter users commonly use hashtags to pinpoint a specific occasion. This enables users to search and participate in a relevant discussion. Mean Sorensen similarity also known as F1 score, [170] was calculated for each of the hashtags. As Sorensen similarity score is within a range of 0 (no overlap) to 1 (identical). Greater values showed greater constancy among a major population of active users.

Occupational Twitter Dataset

While most of the datasets were generated by using relevant keywords/hashtags, the occupation dataset was generated by using the biographical content of each of the target users. The Occupation dataset of Table 2.1 shows the details of the dataset. Occupational Twitter Dataset has public access and maps $5,191$ Twitter users to 9 major occupational classes [149]. The dataset consists of User IDs and the historical tweets of each of the users. The Occupational prediction problem is considered as a multi-classification problem as the model focuses on predicting multiple occupational classes. The occupations of each of the users were manually labeled with the aid of Standard Occupation Classification (SOC) from the UK [2] which is shown in detail in Table 2.2. The dataset was initially used to predict the occupation of the main Twitter user based on their historical tweets [149]. Later, the dataset was further extended, to analyze deeper into the network information. The biographical descriptions of the following and followers' IDs for each of the main user ID [141] were added. Biographical descriptions were extracted from the 160-character-long summary that a user writes about themselves in their profile. Thus, the extended dataset had the followers and followings information for about $4,557$ main users. Due to account suspension and protected tweets, the remaining users account could not be reached. Table 2.2 shows the occupational class distribution of the users' occupational dataset. The biographical information of the main users were not taken into account since the main users' occupations were manually annotated in the dataset.

In order to construct the social network, each follower/following relationship is considered as an undirected edge for predicting the occupational classes of the main users [141]. In this social network, the main Twitter users are considered as being connected with one another

[2]https://www.ons.gov.uk/

Occupational Class	Standard Occupation Classification	Users
1	Managers, Directors, Senior Officials	461
2	Professional Occ.	1,611
3	Associate Profession, Technical Occ.	926
4	Administrative Secretarial Occ.	162
5	Skilled Trades Occ.	768
6	Caring, Leisure, Other Service Occ.	259
7	Sales and Customer Service Occ.	58
8	Process, Plant, Machine Operatives	188
9	Elementary Occ.	124

Table 2.2: The table shows the major groups (left column) and classified jobs with multiple sub-major groups (middle column) by Standard Occupation Classification. The right-most column represents the number of main users [141].

via common followers and following (follow) IDs. The follow IDs which only connect a few of the main IDs are considered to be weak since the flow of information between the main user IDs will be less than these follow IDs. As a result, the network was further refined by keeping only the follow IDs which have more than 10 connections to the main user IDs. After performing the refining step an unweighted graph was constructed in which all the main IDs were connected to each other and the refined graphs consisted of $34,630$ unique users including the 4557 main users. In the network, only 2550 main user IDs have at least one direct connection with another main user ID. Thus, when constructing the network model the main user IDs often shared common follow IDs which enabled the researchers to extract rich network information.

Non-Twitter Based Datasets

Datasets for validating homophilous models were generated from other types of social media platforms as well such as the Weibo Sina website which is a Chinese microblogging site[190].

Similarly, Furthermore, synthetic networks were also developed by using an artificial undirected network. However, the artificial dataset was not thoroughly described in depth [85]. Datasets: 3, 7, and 8 from Table 2.1, shows the datasets generated from other sources. The following paragraphs give an overview of how these datasets were developed for analyzing the effects of homophily.

Asian Telecom Dataset

For studying the effects of homophily for product purchasing [117], purchases of Caller Ring-Back Tones (CRBT) data were provided by an Asian mobile network. This network data was mainly used for predicting consumers' product choice decision and purchase timings. The dataset consisted of three months of detailed 300 million phone call histories of the company's approximately 3.7 million customers. The call attributes were the caller or callee phone numbers and the duration of the phone conversation. The CRBT product was bought by 750,000 customers. The pattern of the CRBT purchasing data was explored to find out the main driving forces of the customers to buy the product. The communication between friends were tracked and analyzed using this dataset. When friends of the customer get exposed to the ringtone by calling them and purchase the same product. Moreover, CRBT is a cheap and economical product that has been purchased by more than $750,000$ customers so the researchers claimed that the dataset of phone call histories provides a convenient platform for studying communications on this product.

Online books dataset

Datasets from amazon.com and amazon.co.uk were used to study the effects of homophily. The data was collected from over 3 million books which included 778,005 British and 1,461,206 American books. Book's ISBN, name of the authors, and literary genre meta-

data were analyzed. The unique ISBN for each of the books was fetched from Amazon's website. In most of the cases, different editions with individual ISBNs were released for the same book title across the two markets(amazon.com and amazon.co.uk).

Weibo Users Dataset

In order to validate the proposed topic model for link prediction [190], Data were extracted from the Sina Weibo[3]. The dataset contains about 2000 verified users, in which each user is represented by a single document. The link information between the pairs of users was also collected, when both the users' posts were present in the dataset. The link information refers to three types of interactions which include mentioning, retweeting, and following in the Weibo website.

Generally, Twitter Search API is used to generate the desired network data for validating the proposed homophilous models. Users and their textual tweets are mainly used as features to develop the dataset. Social media platforms have other features such as images and videos which are posted by users. However, such features have not yet been included in the datasets for evaluating homophily. Furthermore, the size of the datasets varies a lot, ranging from 2000 users' posts to $75,000$ users' posts. Yet, no benchmark has been set to have a minimum standard size of dataset to validate any of the proposed models. Other than using Twitter Search API, microblogs, and artificial data such as Weibo Sina and artificial undirected network has also been used to generate the network data.

[3]https://www.weibo.com

2.7 Conclusion

The homophily principle in the domain of social network analysis is an important concept that has been studied broadly. It has been used to examine the behavior of users on social media platforms. Generally, in social media, users tend to connect with others where they have similar interests with one another. Several studies have been carried out to study the homophilic patterns especially among majority and minority groups. These studies have deduced that majority and larger groups receive information faster than smaller, minority groups. The information reaches like-minded individuals quicker. Besides, multiple types of models have been proposed for measuring the degree of homophily. Network modeling and topic modeling have been mainly used for analyzing the strength of a relationship of a user with the user's nearest neighbors. Network modeling was conducted on network features and topic modeling was conducted on users' textual tweets respectively. Recently, the effect of homophily has also been studied by combining textual and network features with a highly efficient neural network model. However, the content of the interactions occurring between users in social media are still inadequately understood[186]. Furthermore, the content of social media ranges from texts to videos, which needs different types of analysis. Most of the studies focused either on textual posts, hashtags tweeted by users, mentions of users, or users' network connections. However, there is a research gap as comprehensive studies have not been conducted, concerning images and videos posted by users. Whether these contents of social media have any effect on the degree of homophily in online platforms is not fully understood as of yet. Therefore, state-of-art methods should also focus on measuring the level of homophily by using these features.

In this chapter, we presented a survey on the usage of the Homophily principle for computing social network analysis. This thorough survey will enable researchers to explore

new methods to measure the degree of homophily. In summary, multiple methodologies have been proposed over the years to measure the level of homophily by using different types of modeling approaches. This includes topic modeling, network modeling, ERGMs, where each of the models has its own merits and drawbacks. These models are validated by either using synthetic data or by using real-life data from social networking sites such as Twitter. However, the range of data used by each of the proposed models varied to a great extent ranging from posts from only 2000 users to 75,000 users' posts. Therefore, for better comparability of different features and methods, we argue for a benchmark dataset for homophily detection.

Chapters 3 and 4 will describe deep-neural network models combined with word-embedding and a link prediction model for predicting social network homophily of Twitter users based on their bio content. The homophily principle is used as a motivation for designing the link prediction model. Link prediction is calculated based on the similarity between two entities and it can be used to predict future links in social networking platforms [46]. One of the main purposes of this research is to build efficient recommender systems for targeted audiences. For example, it can be used to advertise the journals of the latest medical articles to a network of doctors instead of a network of users working in IT sectors.

Chapter 3

Identifying health related occupations of Twitter Users through word embedding and deep neural networks

All of this chapter was accepted in the following Asia Pacific Bioinformatics Conference (APBC) 2021 which is in press:
This publication summarises my contribution to a broader research project that uses artificial intelligence techniques to predict the occupations of Twitter users in medical fields. We proposed a method of integrating word embedding with state-of-the-art neural network models in a novel way. The textual contents are formatted as dense vectors using word embedding, which are fed as a sequence of vectors into the neural network models. With the support and encouragement of my thesis supervisors, I took the lead in developing this publication. Due to the peer-review process and open-access policy, we chose to publish in the Asia Pacific Bioinformatics Conference, which was ideal for my first publication at the university.

3.1 Introduction

Twitter is a popular social media platform providing micro-blogging service, in which users can share their views and opinions for up to 280 characters long messages known as "tweets". Although, in other social networking sites such as LinkedIn and Facebook, users have the privilege of filling up their personal information in specific fields. Unlike these sites, on Twitter, users can write a public outline about themselves in only 160 characters known as "Bio". Not only tweets, but bios can also help to extract rich linguistic information, and can be extremely helpful for "user profiling". User profiling is an active area of research since it helps to improve product recommendations and service quality. Besides, the prediction of user occupational class is highly vital for user profiling and for predicting users' demographic features. In previous studies, multiple approaches have been proposed for predicting demographic attributes, which include composing multiple features that have been initiated from the text and network information of users for attaining best performances in terms of predicting classification tasks [69, 127, 149, 78, 7, 141]. Users' texts in social media have been studied extensively to mine the features hidden behind the textual data that include Spatio-temporal and social network information [161, 101]. Earlier research has focused on examining users' attributes such as gender, age, and location which showed that their attributes influenced their use of language [39, 28, 155]. The texts of the users enable us to analyze such properties[178, 165, 163]. For example, user profiling can be applied for designing recommender systems for cost-effective targeted advertisements [67]. In other words, it will be more useful to advertise the journals of the latest medical articles to a network of doctors and practicing physicians which are related to their medical fields than to users who might be working in IT sectors.

Based on the biographical content of the users on the Twitter platform, we aim to pre-

dict a user's medical occupational class. Firstly, in our approach, a vector space model has been generated in which each vector represents the unique tokens of each biographical content. The vectors were later fed into Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT), and A lite BERT (ALBERT) neural network models respectively to perform multi-text classification. Our results have demonstrated that by combining neural network models with the Vector Space Model (VSM), our model has performed better in identifying the medical professions from the biographical contents. This task will be directly applicable in analyzing the medical professional trend on Twitter of users following Twitter medical accounts. We have focused on analyzing Twitter medical accounts since we can get a more concentrated network of users belonging to medical fields. Since users working in various types of medical field will be more interested in following such medical accounts than general users. This task can be applied in examining the users belonging to a wide range of medical fields. Health agencies can recruit users working in various medical fields for new job opportunities [166]. For this study, we created a dataset of users following medical accounts, including their biographical content and a label belonging to an occupational class from the "National Occupational Classification"[1] taxonomy. We have designed a dataset that is similar to the dataset that was built by Preoţiuc-Pietro et al.[149].

The rest of the chapter is organized as follows. Section 3.2 presents background on the relevant studies conducted on predicting occupations of Twitter users and the use of deep learning models on different types of text analysis tasks. Section 3.3 introduces our proposed model, Twitter biographical data collection technique, data processing, and implementation methods. Section 3.4 provides the results and discussion. Finally, Section 3.5 draws the chapter to a conclusion.

[1]https://noc.esdc.gc.ca/

3.2 Related Work

To predict the occupations of Twitter users a dataset was constructed by Preoţiuc-Pietro et al., assigning users in 9 hierarchical business job categories [149]. For predicting the user occupational class, word cluster distribution features from each of the users' historical tweets were used. They were able to achieve 50% accuracy in their multi-classification task. Aletras and Chamberlain have built the user's following connections from the users of the occupational dataset. The user embedding was used as an input into their classification model. They have deduced that major job categories may fluctuate consequentially with regions having inequalities of economic development [7]. Pan et al. have used users' network information and their bio description, over their tweets for predicting their occupational group [141]. They have stated that bio description of users along with their network information, when used as features in the neural network graph, has improved the performance of predicting the occupational group compared to using the historical tweets and have attained an accuracy of 61% [92].

The task of identifying the occupation of users is a multi-classification problem as an individual has to be designated to one of many predefined categories which is a classic research area of machine learning [201]. Specifically for classification task, Deep Neural Networks (DNN) have demonstrated to perform better than conventional machine learning models [42, 43, 195, 14]. One such enhanced model of DNN is the Recurrent Neural Network (RNN), which examines each of the words in the corpus and stores the semantics of past words inspected in hidden layers [175]. RNN has proved to be suitable for text classification problems as it can understand contextual information, which is highly required in learning the text semantically [100]. Many works have also proved that RNN based text classifiers perform better as it has achieved higher precision and accuracy scores. For instance, gated

RNN was used for document-level sentiment classification [175] and sequential short-text classification based on RNNs [105]. However, RNN did not perform significantly well in memorizing from distant past texts, which means that RNN cannot remember the text that is at least five words behind the word that is currently being memorized by the model. To resolve this issue, the Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) models based on RNN has shown to be very effective [73, 40]. A forget gate is added to the LSTM model which enables the model to forget the least important words. In contrast to the LSTM model, the GRU model has only two gates, a reset gate, and an update gate. The update gate plays a role that is similar to the forget gate. The gate decides which information is important to remember and adds it to its gate. The reset gate is responsible to decide how much of the past information should the model forget. The LSTM and GRU models are much more efficient since they can learn and remember more texts and solves the issue of memorizing longer texts [58]. LSTM and GRU models have been used extensively in the state of the art deep learning applications, like speech recognition, speech synthesis, and natural language understanding [156, 179]. Also, LSTM neural network was implemented in text classification tasks by combining LSTM and convolutional neural network [203]. Moreover, as the outputs of RNN are mostly based on the previous contexts, the Bi-LSTM model has been introduced [79]. This structure is effective in knowing the future contexts as it allows the networks to know both the backward and forward information about the sequence at each time step. Bidirectional-LSTM runs inputs in two ways, one from the previous texts to upcoming texts. Similarly, one from upcoming contexts to previous contexts. In contrast, to this approach, LSTM is uni-directional and Bi-LSTM has shown to outperform LSTM as Bi-LSTM at any point in time can preserve information from not only the previous contexts but also from the future contexts. BERT is also a bidirectional model just like the Bi-LSTM model and its major component is that it uses a bidirectional transformer language model

while training on textual data [49, 152]. A Transformer is a machine learning model that considers the ordered sequence of the data [151]. BERT uses two pre-training techniques [49].

- Masked language model- Masking is carried out in three different ways. For example, if the sentence to be trained is "My dog is hairy" [49] and the word "hairy" is selected to be the token, then masking is done either by replacing it with a $< Mask >$ token i.e., "My dog is $< Mask >$" or with a random token e.g. "My dog is an apple" or by keeping it as it is i.e., "My dog is hairy" [49]. To get the context of a word these three methods are used together [19, 151].

- Next Sentence Prediction - The model is given a pair of sentences and the model has to predict whether the second sentence comes after the first sentence or not in the corpus [49, 185].

This is the reason why Bidirectional Encoder Representations from Transformers (BERT) is significantly different from the other textual classifiers [49].

As the quantity of training data and the model size increases, the performance of the model increases [102]. However, with the increase in model size, it becomes difficult to pre-train the model because of the Graphical Processing Units (GPU) memory limitations and it takes longer training times [102]. ALBERT was introduced to solve this issue. ALBERT has the same architecture as BERT. However, ALBERT uses two parameter-reduction techniques to significantly decrease the number of training parameters of BERT. They are:

- Factorized embedding parameterization [102], splits down the larger word matrix into smaller matrices so that the complexity is reduced [19].

- Cross-layer parameter sharing stops the parameters from increasing as the depth of the neural network increases [102].

Both the techniques drastically reduce the training time of the model [102].

With the growing importance placed on collecting data, it has now become possible to develop more efficient and trustworthy classification methods to study and analyze biographical data for user profiling tasks. In our research, we propose a method developed on word embedding and neural network models to predict the medical occupations of users from their biographical contents for designing more effective recommender systems.

3.3 Methods

In this study, our task is a multi-class text classification problem where we want to identify the most likely medical profession for a given user based on their biographical content. At first, we have generated a vector space model in which each vector represents the unique tokens of each biographical content. Then the vectors are fed into Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and BERT neural network models respectively. These models then perform multi-classification, identifying the occupation of the user. Our results have shown that neural network models along with Vector Space Model (VSM) have performed well in identifying the medical professions from the biographical contents.

Our Approach

At first, from the unlabeled raw data, a term index Vector Space Model (VSM) is generated, in which the vectors depict the position of the unique tokens in the textual data. Then, a

"Pharmacist, professor and specialist in geriatrics."

Figure 3.1: Example of a biographical content

sequence of vectors representing each bio content is produced using the Vector Space Model (VSM) and given as input into an LSTM, Bi-LSTM, GRU, and Bidirectional Encoder Representations from Transformers (BERT) neural network models respectively that executes a multi-classification task. We, at first, converted our occupation classification into a sequence classification problem, so that the task can be managed appropriately by the neural network classifiers. In Figure 3.1 an example of a biographical content identified of a user belonging to the pharmacist category is shown.

Representation of the biographical data

Our system specializes in feeding raw corpus into the deep learning classifiers. One of the major drawbacks of textual data is, these data cannot be directly represented as dense vectors as the terms from each text tend to have variable lengths. To solve this issue, distributed representations of words (word embedding) are used, where the terms in the text are formatted as dense vectors. This can be done with the aid of padding ("pad"). Padding allows us to get a fixed size input for the vocabulary of all the unique tokens in the corpus. In the vocabulary, the texts cannot be directly portrayed as vectors, and to resolve this issue, the unique tokens are placed with their indices. The biographical texts are converted into a sequence of indices (positive integers), instead of a sequence of tokens. Then, a dense vector is created from the term index sequences, where their representations are aligned to that of the original text. Previous works have also shown that the distributed representation of

words in vector space is embedded with rich semantic and syntactic information [126, 125, 144]. Therefore, we created a vector space model from the term index representation of the texts in which the dense vectors were built from the biographical contents' index-terms. By representing the data in this approach, we expect to feed the classifiers with the crucial knowledge embedded in each of the textual contents.

Text	Pharmacist	professor	and	specialist	in	geriatrics
Tokenized term index	[$i_{Pharmacist}$]	[$i_{professor}$]	[i_{and}]	[$i_{specialist}$]	[i_{in}]	[$i_{geriatrics}$]
Original term index	467	43	254	78	51	165

Table 3.1: Text, Tokenized term indexes

In Table 3.1, the first row shows the sequence of tokens (or term) of the biographic content. The tokenized representation of indices i(term) in the vocabulary of each of the terms is shown in the second row. Lastly, the third row shows the order of the original term indices where the values are determined by the locations of the terms in the vocabulary.

Twitter Biographical Data of Users

As far as we are aware, there are no existing datasets available that provide a convenient way to predict the medical professions of users. Hence, we have built a dataset that maps users to their medical occupations based on their biographical content. Similar to the approach taken by Preoţiuc-Pietro et al., we have examined the Twitter users following Twitter medical accounts having more than 100K followers. The Twitter users who had mentioned their medical profession in the Twitter user description field ("Bio"), those description fields were annotated. The users were selected from 1% of the sample taken through the Twitter

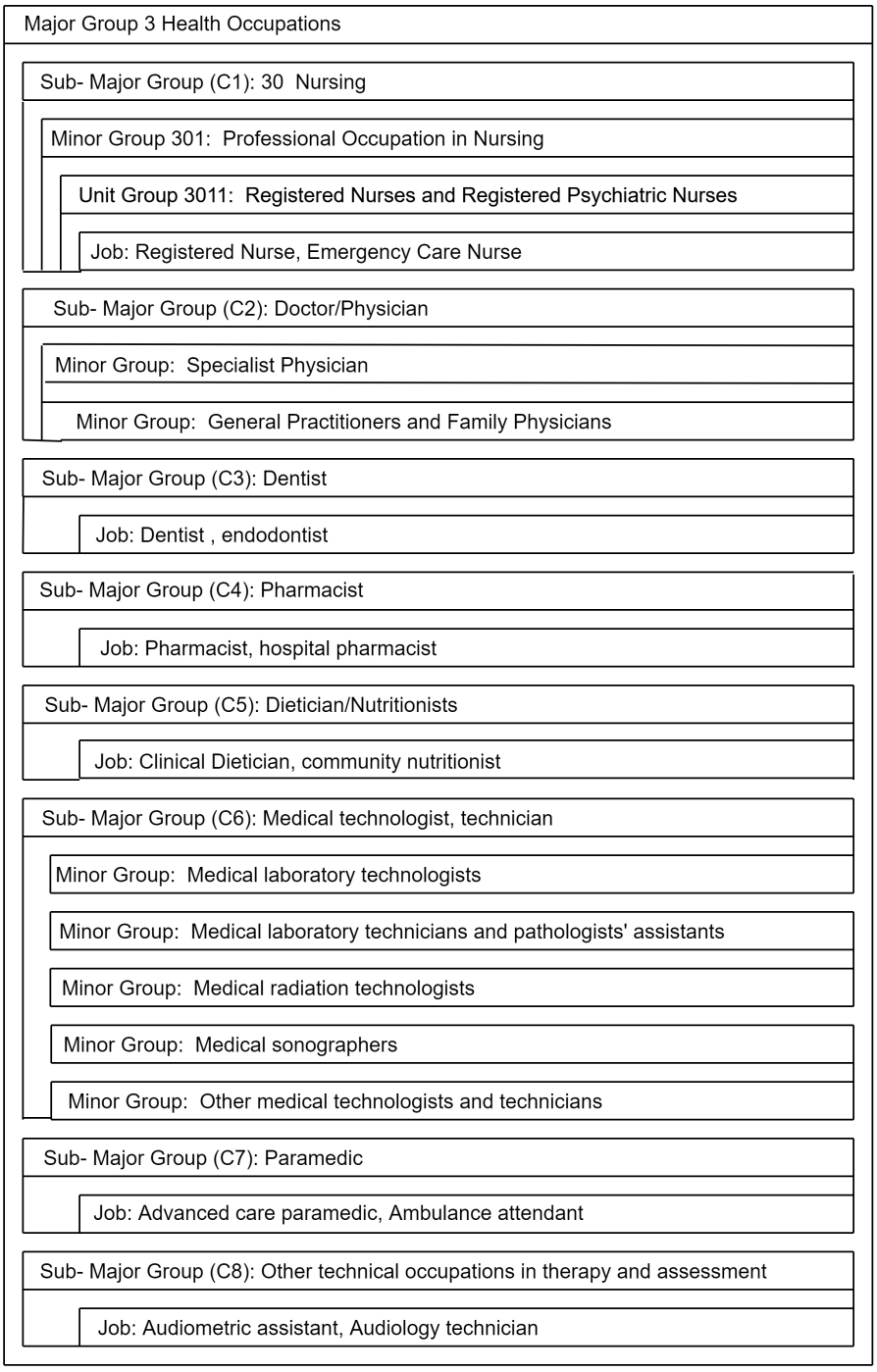


Figure 3.2: Subset of the NOC classification hierarchy

API [124, 33]. By analyzing the description fields of 45,000 users approximately, we found out the following categories: empty description (10.75%), random contents (17.23%), found user information but not related to medical occupations (55.87%), and medical occupations related information (16.15%).

To map the Twitter users to their respective occupations, we have used the standardized job classification taxonomy which is provided by the National Occupational Classification (NOC). NOC is a Canadian government system, which is developed by the Office of National Statistics for categorizing occupations. The NOC scheme consists of seven major medical professional groups coded with digits from 1 to 7. Each major group is split into sub-major groups that are enlisted with 2 digits, in which the first digit represents the major group and the second group represents the sub-major group. Each sub-major group is further categorized into minor groups coded with 3 digits. The minor groups are further broken down into unit groups, which are indicated with 4 digits. The jobs are classified hierarchically based on the skill requirements which are suitable for each of the jobs. The unit groups are the end-leaves of the hierarchy and highlight the appropriate jobs related to each of the major groups. Figure 3.2 shows a part of the NOC hierarchy. Even though there are several other existing hierarchies, we base our research on the NOC classification list since it has been published recently in 2016 and each of the major groups has also added the latest jobs and provided a larger range of job titles, which was highly important in generating our dataset.

For building the dataset, 4-digit NOC unit groups were used to find suitable medical job titles that suited best with the users' medical occupation-related bio contents. As the National Occupational Classification (NOC) unit groups have precise medical job designations. The user accounts were combined into minor (3 digits) categories. To improve the quality of the dataset, the accounts with their bio descriptions were manually examined and the accounts of organizations and companies were removed which had no descriptions related to

Occupational class	National Occupation Classification	Users
1	Nursing Occ.	1691
2	Doctor/Physician Occ.	2137
3	Dentist Occ.	257
4	Pharmacist Occ.	614
5	Dietitian/Nutritionist Occ.	1831
6	Medical technologist, technician Occ.	112
7	Paramedic Occ.	90

Table 3.2: The table shows the number of classes (left column) and classified jobs with multiple sub-major groups (middle column) by National Occupation Classification. The right-most column represents the number of users

minor category medical occupations. Moreover, to ensure that the users were not having any fake accounts, 3-digit minor categories users' accounts that had less than 50 user accounts in their follower list were removed. Finally, a total of 6754 users from the minor categories were combined in 7 sub-major National Occupational Classification (NOC) categories. The seven groups were distributed as such: 25%, 31.6%, 3.8%, 9.1%, 27.1%, 1.7%,1.33%, 0.31% (according to the classes in Figure 3.2) and the number of users for each of the classes are shown in Table 3.2.

To build the vocabulary and VSMs, the preprocessed corpora of unlabeled biographic contents were used for the biographic contents of each of the users. A guideline of annotations was developed to annotate the biographic contents properly. The guideline explains how to identify the occupation of the individual users, and includes descriptions and examples of the medical professions. The bio contents were divided into four portions and four annotators annotated each set of texts separately. Then, the labeled texts were further reviewed by expert annotators. The dataset was divided into the ratio of 70:30 for training and testing size.

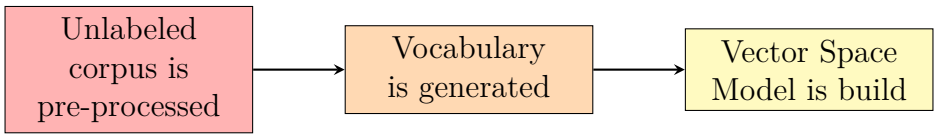


Figure 3.3: The generated vocabulary and VSM from the non-annotated texts

Data Processing

The vocabulary and Vector Space Model (VSM) were created from the non-annotated biographical contents and represented these texts as a sequence of dense vectors. Then, we classified the contents with Long Short Term Memory (LSTM), Bi-LSTM, Gated Recurrent Unit (GRU), Bidirectional Encoder Representations from Transformers (BERT), ALBERT neural networks respectively. At first, a corpus of 6754 unlabeled medical occupation-related bio contents was preprocessed in which URL links and emojis were filtered out. Certain punctuation, duplicates, and texts with URLs were also removed. Secondly, each unique individual index terms were combined to create a vocabulary, and finally, a VSM was generated from the preprocessed textual terms. Figure 3.3 below shows the step by step process of how all unique terms in the vocabulary was generated and how the VSM model was constructed from the vocabulary. The process of representing each biographical content of the annotated corpus into a sequence of index vector is shown in Figure 3.4. In this process, the model at first searches in the vocabulary for the locations of the biographical content terms and then tracks down the required dense vectors from the location information. Finally, the vectors are arranged accordingly for generating a sequence of term index vectors.

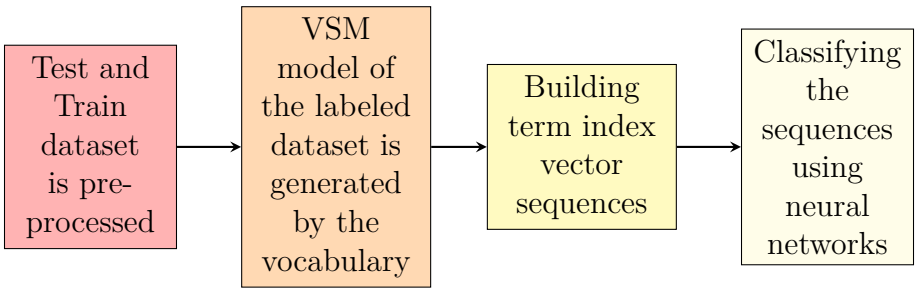


Figure 3.4: Shows the input sequence of vectors into the neural networks for classification. (Figure 3.5 describes the architecture in detail)

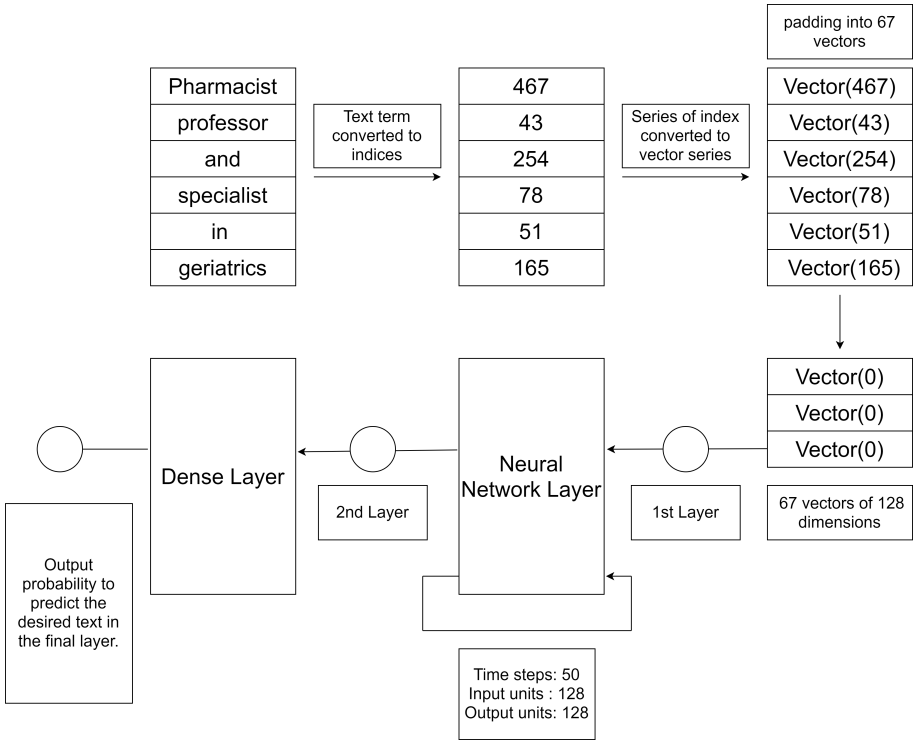


Figure 3.5: Architecture of the Model

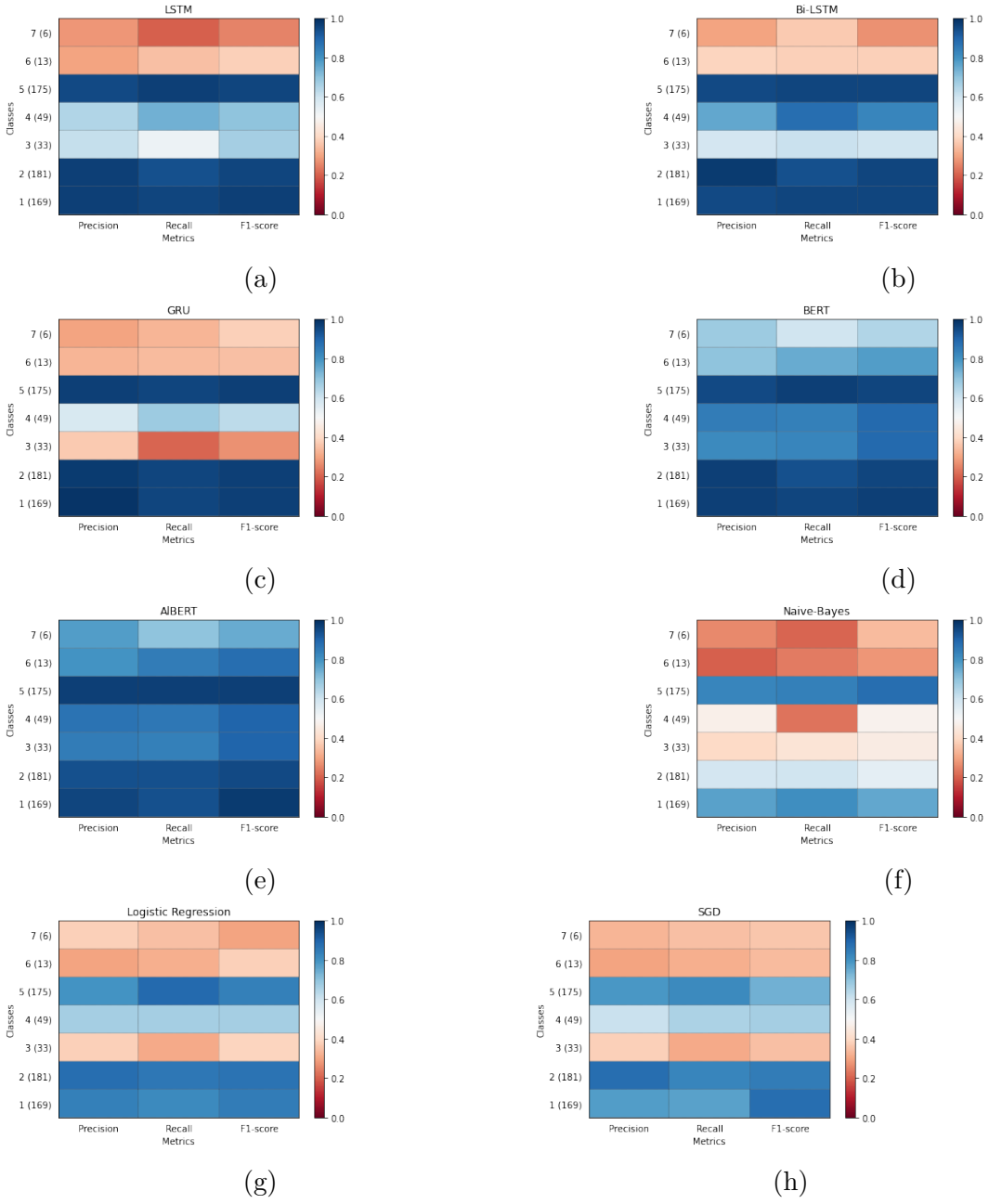


Figure 3.6: Metric Scores of the models

Implementation

The architecture of the neural network model for this study is shown in Figure 3.5. The text of each biographical content is compiled as a sequence of 67 index term vectors as 67 was the largest number of terms collected from each of the contents. If in case, a text was less than 67 index terms, then the sequence was appended with the padding of indices. Each index term was formatted as a 128-dimensional vector of the index corresponding to the term in each of the biographical content. For each text, a sequence of 67 vectors of 128 dimensions was fed to each of the neural network classifiers. The neural network models executed the input sequence and generated the output. Our neural network models were based upon the implementation in Keras, and TensorFlow was used as the backend and high-end Graphics Processing Units (GPU) were used to execute the DNNs. Each of the models was composed of three layers: word embedding layer, Long Short Term Memory (LSTM)/Bidirectional LSTM (BiLSTM)/Gated Recurrent Unit (GRU)/BERT/ALBERT layer, and the final dense layer to process the results. The dense layer is used to process the transition functions generated from the neural network model layer from the input sequences and yields the outputs in n dimensional vectors where n represents the number of classes. Each of the models was trained individually, and accuracy was recorded in each of the epochs. We have observed that the accuracy became stable at around 15 epochs, on average for all the models; so we have selected 15 epochs for training our models. A batch size of 256 and learning rate of $1e^{-5}$ was used for our models. Table 3.3 below further shows the hyperparameters for the pre-trained BERT and ALBERT models.

Hyperparameters	Values
Drop out ratio for attention probability	0.1
Non-linear activation function	relu
Hidden drop out probability	0.1
Embedding size	128
Maximum sequence length	67
Attention heads	12
Hidden layers	768
Hidden groups	1
Inner group number	1
Batch size	256
Learning rate	$1e^{-5}$

Table 3.3: Hyperparameters of pre-trained BERT and ALBERT models

3.4 Results & Discussions

We have calculated the recall, accuracy, and precision weighted scores to evaluate the performance of our approach of combining word embedding and each of the models respectively. As our dataset is highly imbalanced, we have used weighted metric scores as shown in Table 3.4. As the weighted scores consider the label imbalance issue into account. We have used the scikit-learn library for computing the metric scores [145]. The results have shown that ALBERT has performed significantly better compared to the other neural network models. ALBERT has achieved the highest accuracy, precision, recall, and F1-score. Furthermore, we have compared the performance of these models with the traditional machine learning models as well, to validate whether the DNN models perform better than the traditional machine learning models. Overall from the scores, we can see that the DNN model outperformed in predicting the occupational classes of the users. This may be because DNNs try learning high-level features from the textual data. Moreover, due to the availability of high-end, robust computational engines like GPU, it is possible to execute the DNNs. Whereas, for machine learning algorithms, domain expertise is highly required for hardcore feature ex-

traction which makes the task much more challenging. Among the DNN models, ALBERT, BERT, Bi-LSTM, and LSTM have higher precision scores compare to recall scores which suggests that these models predicted more true positives than false positives. Besides, the GRU model has a higher recall value compare to the precision score which shows that GRU can correctly identify more true positives and fewer false negatives. ALBERT took the least amount of time for being trained with the corpus due to its parameter-reduction techniques. In contrast to ALBERT, BERT took the highest amount of time for training. Moreover, as ALBERT takes fewer parameters compared to the BERT model, the training period of ALBERT is less than the amount it took to train BERT. BERT models are already pre-trained with Wikipedia (2.5B words) and BookCorpus dataset (800M words) [102]. As the BERT models are already pre-trained and it is capable of learning complex features faster than the other DNNs.

Model	Accuracy	Precision weighted	Recall weighted	F1 weighted score	Training Time (s)
Naive-Bayes	73.12	0.62	0.60	0.64	586
LR	78.07	0.70	0.70	0.70	574
SGD	79.83	0.74	0.78	0.75	607
GRU	89.76	0.83	0.86	0.88	357
LSTM	90.75	0.89	0.85	0.84	453
Bi-LSTM	91.86	0.86	0.85	0.88	762
BERT	94.78	0.90	0.89	0.89	908
ALBERT	**95.83**	**0.92**	**0.91**	**0.90**	**350**

Table 3.4: Metrics Scores

We have also generated a heat map to analyze the performance of all the classes in

each of the models which are shown in Figure 3.6. The color of the heat map changes from red to blue, from 0.0 to 1.0 respectively where 0.0 represents the lowest score and 1.0 represents the highest score. From Figures 3.6d and 3.6e we can see that Bidirectional Encoder Representations from Transformers (BERT) and ALBERT were able to identify and classify most of the classes compare to the other models with the metric scores ranging from 0.6 to 1.0. Although there were fewer data points for classes 6 and 7, yet it could predict the classes better than the rest of the models. The number of data points for each of the classes is shown in Table 3.2. The Bi-LSTM model was able to classify the classes which have higher data points compared to the classes which have fewer data points shown in Figure 3.6b. It predicted classes 1, 2, and 5 which has a better score for precision, recall, and F1 score. Followed by the Long Short Term Memory (LSTM) model in Figure 3.6a which could predict classes 1,2 5 properly compare to classes 3 and 4 which has fewer data points than the classes 1, 2, and 5 as the intensity of blue color fades in the regions of classes 3 and 4. Similarly, the GRU model in Figure 3.6c, the metric scores shows that the model also performed well for classes 1, 2, and 5 but it performed lower than the LSTM model when predicting the classes of 3 and 4. However, GRU performed slightly better than LSTM and Bi-LSTM in predicting the classes of 6 and 7 which have the least data points. Out of the machine learning models, Naive-Bayes in Figure 3.6f performed the worst as it could neither predict the classes which had higher data points and lower data points properly. As, for classes: 1,2,3,4,6, and 7 the score ranges from 0.0 to only 0.5. Compare to Naive-Bayes model, LR and Stochastic Gradient Descent (SGD) in Figures 3.6g and 3.6h have performed better. Overall, most of the models were able to predict classes, which have higher data points compared to the classes that have lower data points. This suggests that with an increasing number of biographic contents for each of the classes, the performance of the models would increase especially of the Neural Network models. One of the reasons why

most of the models performed with a lower score for classes 6 and 7 is because the job titles vary a lot in these 2 classes as it has several minor groups, and many of the job titles belong to these categories are not that much common.

Based on our observations, we can conclude that the neural network models especially ALBERT perform well as the neural network classifier along with word-embedding may have extracted linguistically rich semantic information embedded in each biographic content to the ALBERT classifier and so this approach results in better classification performance.

3.5 Conclusion

In our approach, we have combined word embedding with deep learning neural network models for predicting the professions of users working in the medical fields. Our research has shown that by composing word embedding with neural network models it has outperformed the conventional machine learning classifiers in identifying the medical occupations of users. Moreover, traditional classification methods depend mostly on features developed by human intelligence which can be very challenging to design. Hence, text classification tasks can tend to become very slow when there is a burden to construct any particular attribute or feature. Thus, classification tasks can be executed much faster when there is no need to design any feature or specific attribute. In the future, the medical occupation dataset can be further extended by adding the network features from each of the users. Network features can be pre-processed from the follower/ following IDs from each of the users, and the features generated from the social network graphs from each of the users can further contribute to improve the occupation prediction of the users.

Chapter 4

Analysis of link prediction using Node2vec with Deep learning framework

All of this chapter is submitted in a peer-reviewed journal :

To continue research efforts on designing efficient recommender systems, we present a novel link prediction method for predicting links in a particular network. The two main contributions here are as follows: a deep learning framework named NODDLE (integration of NOde2vec anD Deep Learning mEthod), where we merge the features extracted by Node2vec and feed them into a four-layer hidden neural network. Our systems NODDLE takes advantage of adaptive learning optimizers for boosting up the performance of link prediction. My responsibilities in this chapter covered from research, development to implementation and article writing.

4.1 Introduction

In social network analysis, link prediction is considered to be a fundamental problem, mainly because of its importance in broad applications of social networks [3, 112, 182, 26, 35]. For example on social media platforms, designing a recommender system is often achieved by link prediction [3]. Moreover, it can be used in cybersecurity inspection to identify credit card fraud, build effective recommender systems for shopping and film suggestions in the e-commerce sector, and locate terrorist groups based on criminal and terrorist activities [35, 147]. In bioinformatics, link prediction is applied in predicting protein-protein interactions containing important information about biomolecular behavior. Such interactions have the potential to reveal answers hidden about diseases and cures [4]. Therefore, predicting such upcoming links is a crucial component of graph mining and has been used extensively in multiple fields.

The main objective of the link prediction problem is to predict the unseen edges that are going to emerge in the graph. Based upon the *snapshot assumption*, when a snapshot of a graph *G(t)* at time t is given, link prediction is assigned to compute which new upcoming links will emerge in the future graph $G(t')$ within the time period $[t, t']$, where $t' = t + n$ (n is the sequence of snapshots) [115]. Link prediction is implemented on real-world network graphs, that are often too massive and tend to be *dynamic* in nature as such graphs are evolving at an extremely high speed. In addition, link prediction uses proximity-based measures, such as the Jaccard coefficient, Resource Allocation, and Adamic Adar metric to measure the probability of the upcoming links to the network [11]. The features extracted are based on the local nodal properties as these functions use the information available only from local proximity of the nodes. Although these metrics are used widely in multiple applications because of their simplicity and interpretability, the problem arises when social

network graphs become large with multiple users. As a result, predicting future links with these measures becomes a very challenging task. Most importantly, hidden and meaningful knowledge lies between the nodes and edges of networks [4]. The analysis of these graphs is an extremely difficult task especially when large-scale network data is composed of billions of nodes and edges [176]. Thus, it is highly important to build effective algorithms for network analysis.

Traditional approaches consisted of using link prediction statically by using only a single snapshot of a network for predicting future links. This prediction task is a time-dependent problem, where a network enlarges over some time [153]. Hence, the dynamic network concept was initiated in which the structure of the network is captured in multiple snapshots over a span of time [202]. Compared to static link prediction, dynamic link prediction is considered to be more valuable and challenging. The evolvement of the network structure offers much more information which not only adds a whole new dimension in the process of network analysis but also helps with achieving a better link prediction performance [154]. The problem arises when the number of edges and nodes increases at a faster rate as it becomes very challenging to extract or infer any reasoning and information from the whole network [65]. Dimensional reduction techniques have been used to solve this issue, which transforms the nodes of a graph into lower dimensional latent representations [157]. These representations can be used as features for executing tasks in graph mining, such as clustering and link prediction [187].

Similarly, network representation learning algorithms such as Node2vec have also been used to tackle this issue, in which node2vec conducts high order proximity by escalating the probability of finding successive neighboring nodes within a fixed length of random walk [65]. This method can efficiently find the equilibrium position between breadth-first search (BFS) and depth-first search (DFS) graphs by developing random biased walks. As a result,

it can succeed in embedding rich quality data, enabling node2vec to preserve the structural balance of the node communities.

Although node2vec has been successful at achieving high link prediction performance, it still has many shortcomings [37]. Firstly, node2vec is a local approach that takes short random walks to get exposed to only the local neighborhood of nodes [36]. Hence, completely it ignores the global relationship of nodes that might have longer distances. Due to taking short walks, the learned representation may be unable to comprehend the important global structure of the model. Secondly, node2vec uses the Stochastic Gradient Descent (SGD) method for resolving non-convex optimization problems, where the non-convex constraints may have various regions and many locally optimal points within each region [8, 81]. The algorithm repeatedly gets updated when SGD is used as the objective function. This causes the optimal points to vibrate frequently and possibly causes them to get stuck in a local minimum. Due to the complexity of the growing networks, recent researchers have focused on applying deep learning techniques to evaluate the complex relationships that exist in graphs and visualize the hidden patterns [181].

In order to tackle these problems we propose NODDLE (integration of **N0**de2vec an**D** **D**eep **L**earning m**E**thod), a deep learning framework in which we combine the features extracted by node2vec algorithm and feed into four layers of hidden neural network and optimize its performance by using different types of optimizers which includes Adaptive Moment Estimation (ADAM), Adamax, An Adaptive Learning Rate Method (ADADELTA), and Adaptive Gradient Algorithm (ADAGRAD) respectively. Next, we have compared our approach with the benchmark methods that include Adamic Adar (AA), Jaccard co-efficient (JC), and Preferential Attachment (PR) [53, 94, 194].

The rest of the chapter is organized as follows. Section 4.2 presents background on the previous studies conducted on link prediction of social networks with heuristic based, machine

learning and deep learning approaches. Section 4.3 introduces our proposed approach in details, including our data preparation method and the method for combining node2Vec with our deep learning framework. Afterwards, in section 4.4, we validated our proposed approach on real-world social network data and analyzed our results. Finally, Section 4.5, draws the conclusion of the chapter.

4.2 Related Work

Heuristic Similarity Metrics

Liben-Nowell and Kleinberg proposed a link prediction problem for social networks using multiple heuristic functions [112]. They found that topological features can be used to predict a future edge between two nodes that showed high "similarity" or "proximity" between the target nodes. Furthermore, their findings conveyed that heuristic measures notable correlation with the predicted future links such as Adamic/Adar and Katz centrality [2, 87].

Many of the researches emphasized on enhancing the performance of heuristic functions by increasing the neighbour-based attributes to second, third or higher adjacency degrees. For instance, Yao et al. presented an improved common neighbors heuristic algorithm that includes nodes with a distance of two hops and used time-decay for recent snapshots to have a greater weight [191]. Kaya et. al. used progressive events to calculate the possibility of future links in a time-weighted fashion [88]. In addition, Deylami and Asadpour proposed a community detection algorithm to identify high activity clusters [50]. Similarity metrics have also been used for common link prediction problems to detect events in social networks and cognitive radio networks [77, 76, 199].

Machine Learning & Deep Learning

Link prediction can be computed by both supervised and unsupervised techniques. Unsupervised methods comprise of developing the heuristic approaches to determine the score for the likelihood of each upcoming link [173]. Similarity metrics are most commonly used for measuring the intensity of the relationship that exists between the nodes. Topological features of the nodes such as common neighbors and graph distances are used to measure the strength of the interaction between the nodes [5]. Conversely, supervised methods involve treating the link prediction problem as a binary classification task in which the edges and non-edges of a network model are employed for training a classifier [107].

Compared to heuristic-based approaches, machine learning techniques have proved to be better at link prediction tasks as these models have received higher prediction accuracies. Yet, the major problem that arises with machine learning models is representing the graphical features, since it is not possible to use the large scale graphs as input into the machine learning models. As a result, researchers have attempted to extract features. For example, Hasan et al. have attempted to extract multiple graph features and implemented the features with various machine learning algorithms such as Decision trees, Naive Bayes, and k-Nearest Neighbors [5]. Similarly, Bechettara et al. have implemented topological-based features of bipartite graphs with decision trees [20]. Doppa et al. proposed a supervised feature vector based approach with k-means classifier for link prediction [52]. Even though machine learning techniques have shown to achieve better prediction accuracy, these methods rely highly on features developed by human intelligence and thus engineering such features is extremely tedious and slow. As a result, most of the state-of-art link prediction techniques utilize deep neural networks for their exceptional learning ability.

A deep neural network model is mainly defined as a group of models in machine learning

consisted of multiple connected layers. The layers generate output-yielding nodes where the parameters of the neural network layers are tuned in continuous iterations to reduce the error between the final output and original value [29]. Li et al. has explored a neural network structure as a conditional temporal Restricted Boltzmann Machine (ctRBM) which expands on the architecture of a RBM to integrate the temporal elements of a dynamic changing network [110]. Furthermore, Zhang et al. suggested the neural network model as a means of feature representation by using the term, Social Pattern and External Attribute Knowledge (SPEAK); these features are used as input in deep neural network models [196]. Ozcan A, has proposed a link prediction algorithm that extracts multi-variable features from heterogeneous networks, and it is based upon non-linear autoregressive neural networks [140]. This method was tested on various datasets and has shown to outperform the existing algorithms which focus on only single variable features. Zhang et al. have proposed a framework that uses graph neural networks for learning general graph features for link prediction [196]. Graph neural networks are defined as a message-passing algorithm, in which the message represents the features extracted from each node in a graph, and their effects on the edges and nodes are learned by neural networks [120]. Their framework has also shown promising results in the online social networking Stanford dataset of Facebook [196]. Therefore, state-of-the-art research has mainly focused on learning multiple features from graphs at an extensive level as such features contain hidden and meaningful insights into link probability. With the rise of complex growing networks, deep learning techniques have shown to produce highly accurate results. Besides, deep learning can model the complex relationships that are hidden in the network data and has the potential to reveal unseen patterns hidden beneath the billions of nodes and edges [153].

Further research is being conducted to improve the performance of link prediction by applying both supervised, unsupervised, and semi-supervised approaches [114, 206]. Semi-

supervised learning is conducted by combining a small proportion of labeled data with a large pile of unlabeled data during the training process. As mentioned earlier, a semi-supervised approach such as node2vec has shown to outperform existing supervised approaches since it can maintain the community structure and can embed better quality information [65]. In addition, neural networks are also being currently used to enhance the performance of link prediction. These novel methods, which take advantage of neural networks, have proven to be effective with high performance [109]. Such methods are capable of producing preferable link prediction results in large complex networks. Even so, a primary disadvantage to such approaches is that the training and prediction process is highly time-consuming.

4.3 Proposed Approach

This section explains the strategy of solving the problem with deep learning method. Algorithms 1 and 2 provide insights on how we have aggregated the connected and unconnected pairs from the network that were used to build the training dataset. The overall steps for preparing graph with connected and unconnected pairs from the raw network graph for generating our model are explained in Algorithm 3. Then we explain the node2vec model for extracting the features from the training network dataset. Finally, we have shown how we have developed the deep neural network framework with improved optimizers for executing AUC scores for the link prediction of the network. Figure 4.1 provides the overview of our proposed approach.

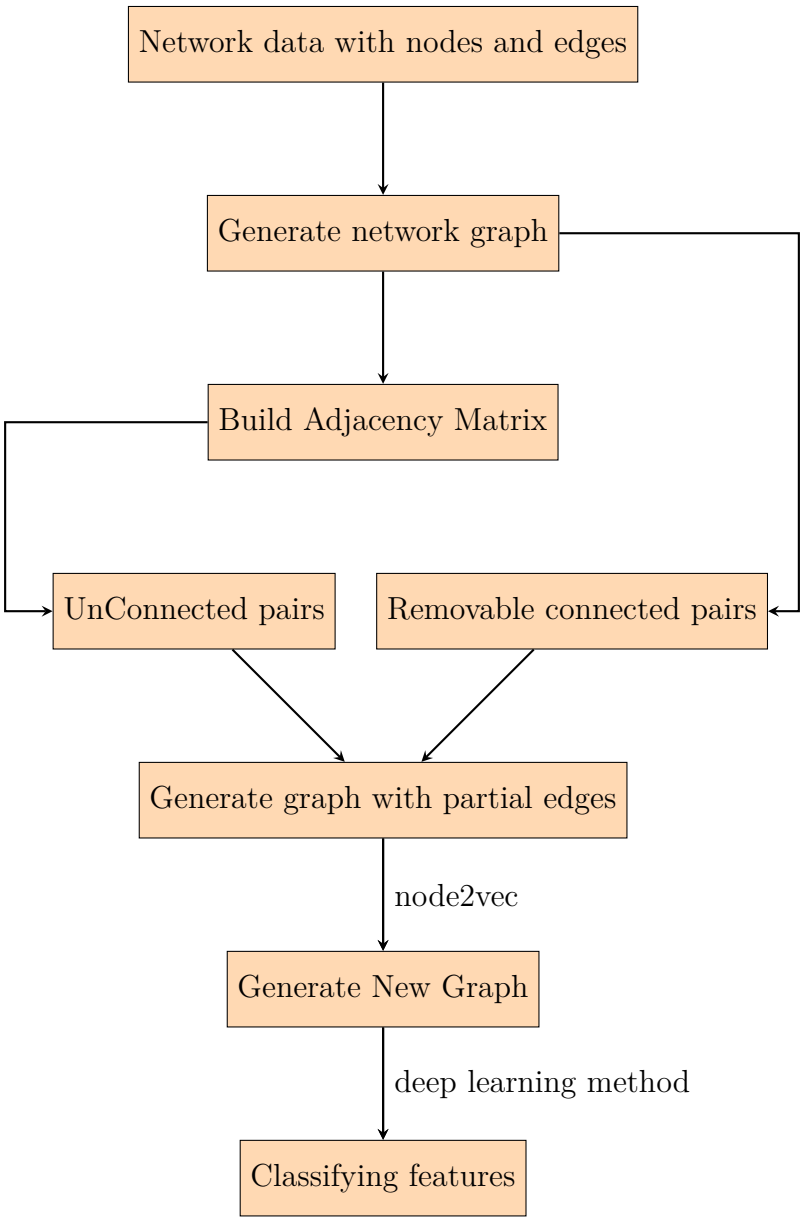
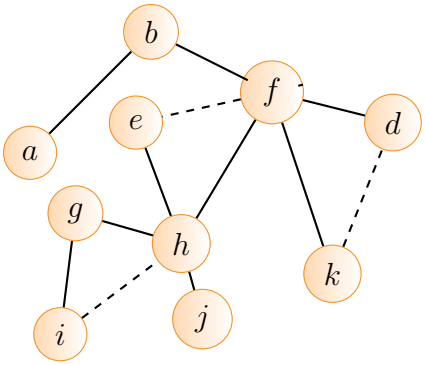


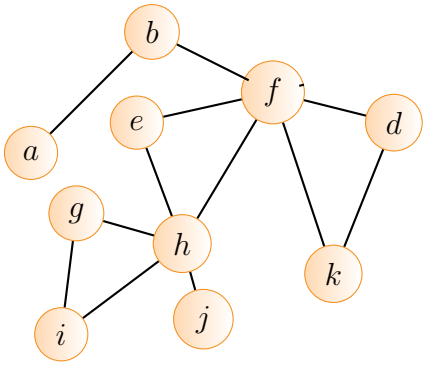
Figure 4.1: Overall structure of the proposed approach

Problem Statement

A sequence of snapshots in time from t to $t+n$ is defined as a dynamic network in which the set of edges in each snapshot depicts the links present at time t. The link prediction problem is such that given snapshots from t to $t+n$, return the score for the possibilities of edges at time $t+n$. Figures 4.2a and 4.2b show a dynamic network with two snapshots. Given the information at time t, we would like to predict the likelihood of link prediction at time $t+n$.



(a) At time t, the dotted lines represents the future links



(b) At time $t+n$, the new links have formed

Figure 4.2: Above figures show a dynamic network with two snapshots

Data Preparation

In real time scenarios, networks data is extremely large, and it is highly imbalanced as it contains a higher number of unconnected nodes than connected nodes. Therefore, it is always a challenge for the model to learn features from connected nodes since the connected node pairs are often much fewer in number than the unconnected nodes. Hence, we provide a way for preparing the data in a computationally economical way that can extract the

unconnected and connected node pairs from large number of imbalanced network data. Below we have discussed in details how we have sampled the positive (connected) and the negative (unconnected) node pairs.

Aggregation of Unconnected Samples

In order to find the negative sample that depicts the unconnected nodes, we have built an adjacency matrix with the aid of *networkx* library. The connected and unconnected nodes are represented as rows and columns for each of the nodes. As the values in the matrix are same for above and below the diagonal of the adjacency matrix, we only focus on finding the positions of the unconnected nodes from above the diagonal to make the approach computationally efficient. Algorithm 1 shows the complete steps for finding the unconnected nodes. We have also used the *networkx* library for finding the shortest path between the unconnected nodes and only select the unconnected nodes that are within the distance 3, after experimentation with other configurations. The unconnected node pairs were labeled as '0', as the unconnected node pairs represent the negative links.

Aggregation of Connected Samples

Some of the edges from the graph will be randomly removed and labelled '1' since, these edges connect the nodes and show the presence of links. Thus, when training the model, it will be able to predict such potential links at time $t+n$. However, it is essential to ensure that the graph's nodes do not become completely isolated when dropping the edges since taking such a step can lead to misrepresentation of data, and the model will be trained poorly. Thus, when removing an edge, we ensure this does not lead to splitting the graph and the number of connected nodes is ≥ 1 . If the removed edge satisfies both of the conditions, only

then the edge is dropped, and the process is repeated for the next pairs of nodes. Algorithm 2 shows the steps for accumulating the positive samples.

Algorithm 1: Finding all Unconnected pairs

Input: Adjacency Matrix as $AdjM$, Graph $G = (N, E)$
Output: All unconnected pairs as UCP

```
UCP ← φ                                   // empty dataframe for all unconnected pairs
for each row in AdjM do
    for each column in AdjM do
        if row.index ≠ column.index then
            if FindShortestLength (G, row, column) ≤ 3 then
                /* using networkx for shortest length function          */
                if AdjM [row, column] == 0 then
                    Append N.row, N.column to UCP
Return UCP
```

Algorithm 2: Finding all Connected pairs

Input: Graph $G = (N, E)$
Output: Connected Pairs as CP

```
CP ← φ                                    // Empty set of connected pairs
for each e in E do
    G' ← RemoveEdge(G) // remove edge from a node pair and generate a
         new graph as G'
    if the new nodes are not completely isolated then
        CP ← Append(G'.node, G'.edge)
Return CP
```

NODDLE (integration of node2vec anD Deep Learning mEthod)

Node2vec algorithm is a feature extraction method, used to generate vector representations of nodes on a graph. It is mainly a local approach that uses random walk to search for the local neighborhood of nodes. The algorithm uses direct encoding and a product-based

Algorithm 3: Data Preparation for generating Model

Input: Node as N and Edge as E
Output: New Graph as G'

```
G ← φ                                                        // Graph
AdjM ← φ                                                     // Adjacency Matrix
UCP ← φ                                                      // Empty set of unconnected pairs
CP ← φ                                                       // Empty set of connected pairs
for each n, e in N, E do
    G ← (n, e)                    // Creating network Graph with networkx library
    AdjM ← AdjacencyMatrix (G) /* Create Adjacency Matrix from nodes and
        edges with networkx                                                    */
    UCP ← UnconnectedPairs (AdjM, G)                         // Algorithm 1
    CP ← ConnectedPairs (G)                                  // Algorithm 2
    G' ←Create NewGraph (UCP, CP)
Return df
```

decoder. Therefore, node2vec embedding is defined as such:

$$DE(s_i, s_j) \cong \frac{e^{z_i^T z_j}}{\sum_{v_k \in V} e^{z_i^T z_k}} \approx (P, R(v_j|v_i)) \tag{4.1}$$

In this Equation 4.1, $DE(s_i, s_j)$ represents the decoded product based proximity value, the probability of visiting to node target node v_j from the source node v_i with fixed length of random walk R is denoted by $(P, R(v_j|v_i))$. $(P, R(v_j|v_i))$ can be calculated for both random and undirected graphs. Cross entropy loss for node2vec is calculated by the following formula:

$$Loss = \sum_{(v_i, v_j) \in Deno} -\log(DE(s_i, s_j)) \tag{4.2}$$

The training set is generated by collecting random walks from a source node v_i in which the N pairs of v_i for each node are collected from the probabilistic distribution of $(v_i, v_j) \sim (P, R(v_j|v_i))$. However, it is extremely expensive to calculate the cross entropy loss because of the high computational costs for evaluating $O(|Deno||V|)$, as $O(|V|)$ has a high time complexity when computing the denominator $Deno$ of Equation 4.1. As a result, node2vec uses various optimization and approximation methods for computing the cross entropy loss

is Equation 4.2 for reducing the computational costs. "Negative sampling" approximation method is used by node2vec to evaluate Equation 4.2. It is easier to calculate entropy loss for labeled data, however, the problem arises when there is a higher proportion of unlabelled data. Contrastive loss has been introduced to calculate the loss for semi-supervised approaches [38]. Contrastive loss calculates the distance between two positive examples and compares it with the distance between two negative examples. Moreover, triplet loss has also been introduced that takes three inputs during training: the Anchor, the positive, and the negative [169]. The Anchor input can be any input, the positive has to be an input belonging to the same class as the anchor, and the negative has to be an input with a different class than the anchor. The Triplet Loss reduces the distance between an anchor and the positive inputs while increasing the distance between the anchor and the negative inputs.

The node2vec takes into account a random set of negative samples for approximately calculating the normalization factor instead of letting the entire set of vertices to be normalized [65]. Additionally, node2vec applies two hyper parameters p and q. The probability of going back to a previous node after visiting a new node is controlled by p. The hyper parameter q controls the possibility to explore new nodes of the graph. When these hyper parameters are employed, node2vec can interpolate between the walks much more smoothly and the approach becomes similar to BFS and DFS. Grover et al. also demonstrated that when the two hyper parameters are well-adjusted, it enables node2vec to preserve the structural balance between the nodes [65]. However, node2vec still has its own drawbacks. node2vec uses SGD method for solving the non-convex optimization problem [104, 61]. The algorithm constantly gets updated when SGD is used as the objective function which causes the optimal points to vibrate frequently, leading the optimal points to dismount into the local minimum range.

Besides, SGD keeps the learning rate constant when the parameters are updated. As a result, SGD cannot adapt the learning rate and adjust it for carrying out greater updates

on lower frequency features [160]. Hence, Adam, Adamax, and Adadelta optimizers have been introduced to resolve this issue. These optimizers are able to incorporate different learning rates with different parameters. Compared to SGD, these optimizers are much more compatible for large network datatsets in high dimensional spaces and most importantly for non-convex optimization objective functions. Furthermore, deep learning techniques are also applied to study the complex relationships that exist with the growing networks. Hence, we are focusing on improving the performance of link prediction by fusing node2vec with deep learning framework, in which the framework is supported with improved optimizers.

Algorithms 4 and 5, show the steps of the node2vec algorithm. The algorithm at first learns the representations of the nodes by generating a random walk with a length of l, which starts from each of the nodes. When the step is taken in each of the walks, sampling is conducted with the transitional probability of θ_{vx}. The transitional probability θ_{vx} of the second order Markov chain is at first calculated so that node sampling can be computed efficiently by using the alias method in $O(\mathrm{l})$ time. In the final phase the transitional probability preprocessing is conducted sequentially and optimization of SGD is used.

Algorithm 4: Node2Vec Algorithm

Input: Graph $G' = (N, E)$, dimension *dim*, Walks per node r, Walk Length l, Context size *h, Return p, In-out q*

Output: final Stochastic gradient descent function as f

```
θ = PreprocessModifiedWeights(G, p, q) G' = (V, E, θ) Initialize walks to empty
for iter 1 to r do
    for all nodes u ∈ V do
        walk = node2vecWalk (G', u, l)
        Append walk to walks
f = Stochastic gradient descent (h, dim, walks)
Return f
```

Algorithm 5: node2vecWalk Algorithm

Input: Graph $G' = (N, E)$, Start node u, Length l, Context size h, Return p, In-out q

Output: $walk$

node2vecWalk(Graph $G' = (N, E)$, Start node u, Length l, Context size h, Return p, In-out q) Initialize $walk$ to $[u]$

$G' = (V, E, \theta)$

Initialize $walks$ to empty

for *walk from* 1 *to* l **do**

  $curr = walk[-1]$

  $V_{current}$ = GetNeighbours($Current, G'$)

  s = AliasSample ($V_{current}, \theta$)

  Append s to $walk$

Return $walk$

Deep learning Framework

In the final step, we built a deep learning network in which the features extracted from node2vec are fed into a four layer hidden neural network. As mentioned earlier, the SGD optimization function of node2vec has limited capabilities to adapt with different learning rates. As a result, for boosting up the performance of link prediction task, we built the deep learning model with adaptive learning rate optimizers: Adam, Adamax, Adadelta, and Adagrad respectively. This approach is treated as a supervised classification problem, where the aim of the network is to yield a single value representing the probability for a given edge. Thus, we end the deep learning framework with a sigmoid activation function, so that the score is between 0 and 1.

Optimizers

Below we have discussed the different types of optimizers that were used with the deep learning framework.

- Adagrad: Adaptive gradient, or AdaGrad, divides the learning rate by the square root of v, which is mainly the cumulative sum of current and past squared gradients up to time t [53]. Moreover, the gradient component is unchanged just like in SGD. The Adagrad is defined as such:

$$w_{t+1} = w_t - \frac{\rho}{\sqrt{v_t + \epsilon}} \cdot \frac{\partial L}{\partial w_t} \tag{4.3}$$

  In Equation 4.3, w_t is the current weight at time step t that needs to be updated, ρ represents the learning rate and $\frac{\partial L}{\partial w_t}$ denotes the gradient descent to update the weight at w_t and ϵ is a constant value.

- Adadelta: Adadelta is a much more powerful extension of Adagrad that emphasizes on the learning rate component [194]. The optimizer is based on updating gradient using sliding window technique instead of aggregating all the previous gradients. In Adadelta, the difference between the current weight and the updated weight is denoted by 'delta'. Furthermore, the learning rate parameter is replaced by T, the exponential moving average of squared deltas and is defined in Equation 4.4.

$$w_{t+1} = w_t - \frac{\sqrt{T_{t-1} + \epsilon}}{\sqrt{v_t + \epsilon}} \cdot \frac{\partial L}{\partial w_t} \tag{4.4}$$

- Adam: Adaptive moment estimation, or Adam focuses on the gradient component by using $\hat{s}$, which estimates the exponential average of the moving gradients [94]. In addition, the learning rate component is calculated by dividing the learning rate ρ by square root of v which is the exponential moving average of squared gradients. The equation is defined as below:

$$w_{t+1} = w_t - \frac{\sqrt{T_{t-1} + \epsilon}}{\sqrt{v_t} + \epsilon} \cdot \hat{s}_t \tag{4.5}$$

- Adamax: AdaMax is a variation of the Adam optimiser which uses infinity norms [94]. The infinity norm is used to calculate the absolute values of the v components in a vector space ('max')and $\hat{s}$ refers to the estimated value of the exponential average of moving gradients, and v is the exponential moving average of previous p-norm of gradients, that is approximately the max function as defined below:

$$w_{t+1} = w_t - \frac{\rho}{v_t} \cdot \hat{s}_t \tag{4.6}$$

4.4 Experimental Results & Discussions

This section will evaluate our proposed model on real-world data network datasets and examine how our model is more effective than the existing benchmark methods, including Adamic Adar, Preferential Attachment and Jaccard Coefficient.

Datasets

We evaluate our model on Facebook[1] and Twitter[2] datasets that consists of nodes and edges. Table 4.1 shows the overview of the five network datasets. First four of these datasets were collected from the SNAP website. Also, with the aid of Twitter API we have extended a Twitter dataset of around 7,000 users who have follow Twitter medical accounts [91]. The extended dataset contains the follower and following IDs of the users working in medical profession. Public biographical contents of the users were used for finding the occupation of the users. The dataset is described in details in chapter 3.3.

[1]https://snap.stanford.edu/data/egonets-Facebook.html
[2]https://snap.stanford.edu/data/egonets-Twitter.html

Dataset	Number of nodes	Number of edges
Twitter	81,306	1,768,149
Facebook1	4,039	88,234
Facebook2	1,046	27,794
Facebook3	546	5,360
Occupation	6,754	470,168

Table 4.1: Details of the Datasets

Results & Discussions

Our proposed model was implemented in Python 2.8.6 and the experiment was conducted on Lakehead University's HPC (High Configuration GPU enabled PC). In our model we have used four layer fully connected deep neural network with 1024 ReLU neurons in each of the hidden layers. Then, developed our deep learning framework using Adagrad, Adadelta, Adam and Adamax optimizers to improve the performance of link prediction task.

We have calculated the Area Under ROC Curve (AUC) scores to evaluate the performance of our approach of combining node2vec and deep learning framework with each of the optimizers, respectively. The AUC score is defined in Equation 4.7:

$$AUC = \frac{D_0 - n_0(n_0+1)/2}{n_0 n_1} \tag{4.7}$$

In Equation 4.7, n_0 and n_1 denotes the number of positive and negative class links, respectively and $D_0 = \sum r_i$, where r_i represents the rank of the index i in the positive class link in terms of similarity index. Also, $AUC \in [0, 1]$, in which the higher the value of AUC, the higher the link prediction accuracy of the algorithm. We have compared the performance of our approach with the traditional link prediction benchmark methods: Adamic Adar (AA), Jaccard Co-efficient (JC) and Preferential Attachment (PA). In AA, the association between two neighbouring nodes with a smaller degree may occur more than a node with a higher

degree [204]. For instance, it will be more likely for two fans of a celebrity to not know each other. Yet, if two users follow a celebrity who has less fans, then those two users will have a higher chance to have similar interests or tend to be in the same social circle. JC believes that the probability of the presence of links is proportional to the number of two nodes' neighbors [66]. For example, if any two Twitter users tend to have similar interests then they have a higher chance of having some type of connections with each other. Research have shown that the rate of an edge to be connected to a node is proportional to the degree of the node [1]. Thus, PA states that the chances of a new edge to be connected to a node is related to the degree of the node such as two popular celebrities will have a higher chance to know each other since they have higher degree compare to two ordinary persons. The equations of the following algorithms are stated in Equations 4.8, 4.9, and 4.10:

$$s_{AA} = \sum_{x \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log k_x} \tag{4.8}$$

$$s_{JC} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} \tag{4.9}$$

$$s_{PA} = k_x \cdot k_y \tag{4.10}$$

Table 4.2 shows the AUC scores we have obtained from the link prediction algorithms. Overall, node2vec and the node2vec optimized algorithm (Node2Vec+DL) has performed better than the traditional benchmark methods. This might be because node2vec algorithms can learn high level features from the network data [65]. Moreover, as high-end robust computational engines like GPU are readily available now, it is possible to execute the DL models. Whereas, predicting future links from large network data is challenging for the existing benchmark methods. Among the DL models, node2vec with Adamax optimizer has received the highest AUC score in Occupation and Facebook1 and Facebook2 datasets. This is because the Adam optimizer is an upgraded optimizer that has the combined features

Link Prediction Algorithm	AUC Score				
	Twitter	Occupation	Facebook1	Facebook2	Facebook3
Node2vec	0.895	0.876	0.938	0.873	0.861
Node2Vec + DL (Adam)	0.902	0.931	0.934	0.862	0.855
Node2Vec + DL (Adamax)	0.916	**0.945**	**0.941**	**0.879**	0.882
Node2Vec + DL (Adagrad)	0.911	0.911	0.932	0.845	0.851
Node2Vec + DL (Adadelta)	**0.924**	0.932	0.908	0.871	**0.863**
Adamic Adar	0.897	0.711	0.898	0.878	0.734
Jaccard-Coefficient	0.897	0.748	0.901	0.856	0.699
Preferential Attachment	0.891	0.803	0.835	0.801	0.76

Table 4.2: AUC Scores of the Link Prediction Algorithms

of RMSprop and Momentum optimizer [184, 41]. The momentum optimizer has shown to be very efficient as it replaces the current gradient with an aggregation of gradients. This aggregation is the exponential moving average of current and past gradients causing the process to be optimized at a faster rate. Moreover, the RMSprop optimizer also calculates the exponential moving average of the gradients instead of taking the cumulative sum of squared gradients. Therefore, enhancing the overall performance. The node2vec with Adadelta optimizer has performed best in Twitter and Facebook3 dataset. The model with Adamax optimizer has performed better than the rest of the optimizers across three datasets which proves that the Adamax optimizer modified over Adam optimizer, performs better than the Adam optimizer. Similarly, the model with Adadelta optimizer has performed better for Twitter and Facebook3 datasets than the Adagrad optimizer. This has demonstrated that the Adadelta optimizer which is a improved version of Adagrad optimizer has achieved better performance score than the Adagrad optimizer. Thus, from the results in Table 4.2, we can see that optimizers of the DL framework have increased the performance of the node2vec algorithm. The model proposed in this chapter has acquired higher AUC scores than the existing benchmark and node2vec method. Also, the AUC scores of the node2vec with

improved optimizers of the DL framework are highest across all the datasets.

4.5 Conclusion

In this chapter, we explored the drawbacks of the node2vec algorithm when boosting up non-convex functions. In other words, the likelihood of falling into a local minimum due to lack of network knowledge and SGD optimizer's incapabilities to execute adaptive adjustment of the learning rate. Hence, such a scenario makes it extremely difficult for node2vec to process sparse social networks. As a result, in this chapter we proposed NODDLE, a deep learning framework, where we have merged the features aggregated by the node2vec algorithm and used them as inputs into a multi-layer neural network optimizing its performance by using different types of improved optimizers such as Adam, Adamax, Adadelta, and Adagrad. Compared to the various baselines, the results of experiments on real-world social networks proved that our approach not only enhances the prediction accuracy, but it is much more effective and efficient.

Chapter 5

Conclusion

Firstly, a survey is presented that focuses on the idea of homophily as well as relevant social network issues Chapter 2. This survey summarises the state-of-the-art methods that have been proposed in recent years to define and quantify the impact of homophily in various types of social networks. This enables us to identify unsolved problems and research gaps. A model is designed based on the textual properties of the Twitter users' bio contents to identify Twitter users working in medical professional fields in Chapter 3. The proposed model composes word embedding with neural network models, that include: LSTM, Bidirectional LSTM, GRU, BERT, and ALBERT. It is observed that by combining word embedding with neural network models eliminates the need to develop any specific attribute or feature.

Lastly, based on the homophily concept in Chapter 4, a link prediction model is proposed by using the Twitter users' followers and following IDs. Link prediction algorithm such as Node2vec yields good result when predicting links in evolving networks. However, the Node2vec's Stochastic Gradient Descent (SGD) method is prone to falling into a local optimum, and as a consequence Node2vec fails to capture the network's global structure. To resolve this problem, NODDLE is proposed. NODDLE is a deep learning architecture that

combines the Node2vec's extracted features and feeds them into a four-layer hidden neural network. Adaptive learning optimizers are used in our approach to improve the efficiency of link prediction.

To summarize, this research work aimed to encourage further comprehensive study of social network sciences by using both textual and network properties of social media platforms with deep learning techniques. This is highly important, as deep learning techniques can process data at higher volume and speed. This thesis can act as a basis by providing solutions to building effective recommender systems for social networks.

Bibliography

[1] Alireza Abbasi, Liaquat Hossain, and Loet Leydesdorff. “Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks”. In: *Journal of Informetrics* 6.3 (2012), pp. 403–412.

[2] Lada A Adamic and Eytan Adar. “Friends and neighbors on the web”. In: *Social networks* 25.3 (2003), pp. 211–230.

[3] Luca Maria Aiello et al. “Friendship prediction and homophily in social media”. In: *ACM Transactions on the Web (TWEB)* 6.2 (2012), pp. 1–33.

[4] Edoardo M Airoldi et al. “Mixed membership stochastic blockmodels”. In: *Journal of machine learning research* 9.Sep (2008), pp. 1981–2014.

[5] Mohammad Al Hasan et al. “Link prediction using supervised learning”. In: *SDM06: workshop on link analysis, counter-terrorism and security*. Vol. 30. 2006, pp. 798–805.

[6] Yahya Albalawi, Nikola S Nikolov, and Jim Buckley. “Trustworthy health-related tweets on social media in Saudi Arabia: tweet metadata analysis”. In: *Journal of medical Internet research* 21.10 (2019), e14731.

[7] Nikolaos Aletras and Benjamin Paul Chamberlain. “Predicting twitter user socioeconomic attributes with network and language information”. In: *Proceedings of the 29th on Hypertext and Social Media*. 2018, pp. 20–24.

[8] Shun-ichi Amari. "Backpropagation and stochastic gradient descent method". In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[9] Jisun An and Ingmar Weber. "# greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter". In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[10] Rajkumar Arun et al. "On finding the natural number of topics with latent dirichlet allocation: Some observations". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2010, pp. 391–402.

[11] Jibouni Ayoub et al. "Accurate link prediction method based on path length between a pair of unlinked nodes and their degree". In: *Social Network Analysis and Mining* 10.1 (2020), p. 9.

[12] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[13] Michael J Barone, Anthony D Miyazaki, and Kimberly A Taylor. "The influence of cause-related marketing on consumer choice: does one good turn deserve another?" In: *Journal of the academy of marketing Science* 28.2 (2000), pp. 248–262.

[14] Tanvi Barot, Gautam Srivastava, and Vijay Mago. "Determining Sufficient Volume of Data for Analysis with Statistical Framework". In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, Cham. 2020, pp. 770–781.

[15] Jose E Barreto and Curtis L Whitehair. "Social media and web presence for patients and professionals: evolving trends and implications for practice". In: *PM&R* 9.5 (2017), S98–S105.

[16] Nikita Basov. "The ambivalence of cultural homophily: Field positions, semantic similarities, and social network ties in creative collectives". In: *Poetics* (2019).

[17] Frank M Bass, Trichy V Krishnan, and Dipak C Jain. "Why the Bass model fits without decision variables". In: *Marketing science* 13.3 (1994), pp. 203–223.

[18] Mark Belford, Brian Mac Namee, and Derek Greene. "Stability of topic modeling via matrix factorization". In: *Expert Systems with Applications* 91 (2018), pp. 159–169.

[19] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text". In: *arXiv preprint arXiv:1903.10676* (2019).

[20] Nesserine Benchettara, Rushed Kanawati, and Celine Rouveirol. "Supervised machine learning applied to link prediction in bipartite social networks". In: *2010 International Conference on Advances in Social Networks Analysis and Mining*. IEEE. 2010, pp. 326–330.

[21] Sven van den Beukel, Simon H Goos, and Jan Treur. "An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle". In: *Neurocomputing* 338 (2019), pp. 361–371.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[23] Jeffrey Layne Blevins et al. "Tweeting for social justice in# Ferguson: Affective discourse in Twitter hashtags". In: *new media & society* 21.7 (2019), pp. 1636–1653.

[24] Sidsel Boldsen, Manex Agirrezabal, and Patrizia Paggio. "Identifying Temporal Trends Based on Perplexity and Clustering: Are We Looking at Language Change?" In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. 2019, pp. 86–91.

[25] Andrei Boutyline and Robb Willer. "The social structure of political echo chambers: Variation in ideological homophily in online networks". In: *Political Psychology* 38.3 (2017), pp. 551–569.

[26] Marco Bressan et al. "Counting graphlets: Space vs time". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017, pp. 557–566.

[27] Doina Bucur. "Gender homophily in online book networks". In: *Information sciences* 481 (2019), pp. 229–243.

[28] John D Burger et al. "Discriminating gender on Twitter". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 1301–1309.

[29] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. "An analysis of deep neural network models for practical applications". In: *arXiv preprint arXiv:1605.07678* (2016).

[30] Juan Cao et al. "A density-based method for adaptive LDA model selection". In: *Neurocomputing* 72.7-9 (2009), pp. 1775–1781.

[31] Dražen Cepić and Željka Tonković. "How social ties transcend class boundaries? Network variability as tool for exploring occupational homophily". In: *Social Networks* 62 (2020), pp. 33–42.

[32] Pierfrancesco Cervellini, Angelo Garangau Menezes, and Vijay Kumar Mago. "Finding trendsetters on yelp dataset". In: *2016 IEEE symposium series on computational intelligence (SSCI)*. IEEE. 2016, pp. 1–7.

[33] Dhivya Chandrasekaran and Vijay Mago. "Evolution of Semantic Similarity—A Survey". In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–37.

[34] Jonathan Chang et al. “Reading tea leaves: How humans interpret topic models”. In: *Advances in neural information processing systems*. 2009, pp. 288–296.

[35] Haochen Chen et al. “Harp: Hierarchical representation learning for networks”. In: *arXiv preprint arXiv:1706.07845* (2017).

[36] Haochen Chen et al. “Harp: Hierarchical representation learning for networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[37] Jinyin Chen et al. “N2VSCDNNR: A local recommender system based on node2vec and rich information network”. In: *IEEE Transactions on Computational Social Systems* 6.3 (2019), pp. 456–466.

[38] Ting Chen et al. “Big self-supervised models are strong semi-supervised learners”. In: *arXiv preprint arXiv:2006.10029* (2020).

[39] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. “You are where you tweet: a content-based approach to geo-locating twitter users”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 759–768.

[40] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

[41] Dami Choi et al. “On empirical comparisons of optimizers for deep learning”. In: *arXiv preprint arXiv:1910.05446* (2019).

[42] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. “Multi-column deep neural networks for image classification”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3642–3649.

[43] Dan CireşAn et al. “Multi-column deep neural network for traffic sign classification”. In: *Neural networks* 32 (2012), pp. 333–338.

[44] Andrea Fronzetti Colladon and Peter A Gloor. “Measuring the impact of spammers on e-mail and Twitter networks”. In: *International Journal of Information Management* 48 (2019), pp. 254–262.

[45] Alexis Conneau and Guillaume Lample. “Cross-lingual Language Model Pretraining”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 7057–7067.

[46] Sergio Currarini, Jesse Matheson, and Fernando Vega-Redondo. “A simple model of homophily in social networks”. In: *European Economic Review* 90 (2016), pp. 18–39.

[47] Morteza Dehghani et al. “Purity homophily in social networks.” In: *Journal of Experimental Psychology: General* 145.3 (2016), p. 366.

[48] Romain Deveaud, Eric SanJuan, and Patrice Bellot. “Accurate and effective latent concept modeling for ad hoc information retrieval”. In: *Document numérique* 17.1 (2014), pp. 61–84.

[49] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[50] Hasti Akbari Deylami and Masoud Asadpour. “Link prediction in social networks using hierarchical community detection”. In: *2015 7th Conference on Information and Knowledge Technology (IKT)*. IEEE. 2015, pp. 1–5.

[51] Ersin Dincelli, Yuan Hong, and Nic DePaula. “Information diffusion and opinion change during the gezi park protests: Homophily or social influence?” In: *Proceedings of the Association for Information Science and Technology* 53.1 (2016), pp. 1–5.

[52] Janardhan Rao Doppa et al. "Learning algorithms for link prediction based on chance constraints". In: *Joint european conference on machine learning and knowledge discovery in databases*. Springer. 2010, pp. 344–360.

[53] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[54] Hirotaka Ejima, Joseph J Richardson, and Frank Caruso. "Metal-phenolic networks as a versatile platform to engineer nanomaterials and biointerfaces". In: *Nano Today* 12 (2017), pp. 136–148.

[55] César G Escobar-Viera et al. "For better or for worse? A systematic review of the evidence on social media use and depression among lesbian, gay, and bisexual minorities". In: *JMIR mental health* 5.3 (2018), e10496.

[56] Daschel Franz et al. "Using Facebook for Qualitative Research: A Brief Primer". In: *Journal of medical Internet research* 21.8 (2019), e13544.

[57] Pablo Gamallo, José Ramom Pichel Campos, and Inaki Alegria. "A perplexity-based method for similar languages discrimination". In: *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*. 2017, pp. 109–114.

[58] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM". In: (1999).

[59] Jannath Ghaznavi and Laramie D Taylor. "Bones, body parts, and sex appeal: An analysis of# thinspiration images on popular social media". In: *Body image* 14 (2015), pp. 54–61.

[60] Eric Gilbert and Karrie Karahalios. "Predicting tie strength with social media". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2009, pp. 211–220.

[61] Yoav Goldberg and Omer Levy. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: *arXiv preprint arXiv:1402.3722* (2014).

[62] Sandra Gonzalez-Bailon. "Opening the black box of link formation: Social factors underlying the structure of the web". In: *Social Networks* 31.4 (2009), pp. 271–280.

[63] Palash Goyal and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey". In: *Knowledge-Based Systems* 151 (2018), pp. 78–94.

[64] Matthew K Grace. "Friend or frenemy? Experiential homophily and educational track attrition among premedical students". In: *Social Science & Medicine* 212 (2018), pp. 33–42.

[65] Aditya Grover and Jure Leskovec. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.

[66] Anand Kumar Gupta and Neetu Sardana. "Significance of clustering coefficient over jaccard index". In: *2015 Eighth International Conference on Contemporary Computing (IC3)*. IEEE. 2015, pp. 463–466.

[67] Ido Guy et al. "Social media recommendation based on people and tags". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010, pp. 194–201.

[68] Yosh Halberstam and Brian Knight. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter". In: *Journal of public economics* 143 (2016), pp. 73–88.

[69] Bo Han, Paul Cook, and Timothy Baldwin. "A stacking-based approach to twitter user geolocation prediction". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2013, pp. 7–12.

[70] Shuangshuang Han et al. "Analyze users' online shopping behavior using interconnected online interest-product network". In: *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2018, pp. 1–6.

[71] Lydia Hanks, Nathan Line, and Wan Yang. "Status seeking and perceived similarity: a consideration of homophily in the social servicescape". In: *International Journal of Hospitality Management* 60 (2017), pp. 123–132.

[72] Itai Himelboim et al. "Valence-based homophily on Twitter: Network analysis of emotions and political talk in the 2012 presidential election". In: *New media & society* 18.7 (2016), pp. 1382–1400.

[73] Sepp Hochreiter. "JA1 4 rgen Schmidhuber (1997)."Long Short-Term Memory"". In: *Neural Computation* 9.8 ().

[74] Tad Hogg et al. "Multiple Relationship Types in Online Communities and Social Networks." In: *AAAI Spring Symposium: Social Information Processing*. 2008, pp. 30–35.

[75] Roger A Horn. "The hadamard product". In: *Proc. Symp. Appl. Math*. Vol. 40. 1990, pp. 87–169.

[76] Wenbin Hu et al. “An event detection method for social networks based on hybrid link prediction and quantum swarm intelligent”. In: *World Wide Web* 20.4 (2017), pp. 775–795.

[77] Wenbin Hu et al. *RETRACTED: An event detection method for social networks based on link prediction*. 2017.

[78] Yanxiang Huang et al. “A multi-source integration framework for user occupation inference in social media systems”. In: *World Wide Web* 18.5 (2015), pp. 1247–1267.

[79] Zhiheng Huang, Wei Xu, and Kai Yu. “Bidirectional LSTM-CRF models for sequence tagging”. In: *arXiv preprint arXiv:1508.01991* (2015).

[80] Gregory A Huber and Neil Malhotra. “Political homophily in social relationships: Evidence from online dating behavior”. In: *The Journal of Politics* 79.1 (2017), pp. 269–283.

[81] Prateek Jain and Purushottam Kar. “Non-convex optimization for machine learning”. In: *arXiv preprint arXiv:1712.07897* (2017).

[82] S Mo Jang and P Sol Hart. “Polarized frames on “climate change” and “global warming” across countries and states: Evidence from Twitter big data”. In: *Global Environmental Change* 32 (2015), pp. 11–17.

[83] Yuping Jin. “Development of word cloud generator software based on python”. In: *Procedia engineering* 174 (2017), pp. 788–792.

[84] Andreas Kamilaris and Francesc X Prenafeta-Boldú. “Deep learning in agriculture: A survey”. In: *Computers and electronics in agriculture* 147 (2018), pp. 70–90.

[85] Fariba Karimi et al. “Homophily influences ranking of minorities in social networks”. In: *Scientific reports* 8.1 (2018), pp. 1–12.

[86] Eva Kassens-Noor, Joshua Vertalka, and Mark Wilson. "Good Games, bad host? Using big data to measure public attention and imagery of the Olympic Games". In: *Cities* 90 (2019), pp. 229–236.

[87] Leo Katz. "A new status index derived from sociometric analysis". In: *Psychometrika* 18.1 (1953), pp. 39–43.

[88] Mehmet Kaya et al. "Unsupervised link prediction based on time frames in weighted–directed citation networks". In: *Trends in Social Network Analysis*. Springer, 2017, pp. 189–205.

[89] Willemien Kets and Alvaro Sandroni. "A belief-based theory of homophily". In: *Games and Economic Behavior* 115 (2019), pp. 410–435.

[90] M Laeeq Khan. "Social media engagement: What motivates user participation and consumption on YouTube?" In: *Computers in Human Behavior* 66 (2017), pp. 236–247.

[91] Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. "Identifying health related occupations of Twitter Users through word embedding and deep neural networks". In: *Proceedings of The 19th Asia Pacific Bioinformatics Conference*. APBC '21 Accepted, In press (2021).

[92] Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. "The Homophily Principle in Social Network Analysis". In: *arXiv preprint arXiv:2008.10383* (2020).

[93] Kibae Kim and Jörn Altmann. "Effect of homophily on network formation". In: *Communications in Nonlinear Science and Numerical Simulation* 44 (2017), pp. 482–494.

[94] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[95] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[96] Ilkka Koiranen et al. "Shared contexts, shared background, shared values–Homophily in Finnish parliament members' social networks on Twitter". In: *Telematics and Informatics* 36 (2019), pp. 117–131.

[97] Hyeokkoo Eric Kwon, Wonseok Oh, and Taekyung Kim. "Platform structures, homing preferences, and homophilous propensities in online social networks". In: *Journal of Management Information Systems* 34.3 (2017), pp. 768–802.

[98] Riadh Ladhari, Elodie Massa, and Hamida Skandrani. "YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise". In: *Journal of Retailing and Consumer Services* 54 (2020), p. 102027.

[99] Mirko Lai et al. "Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter". In: *Data & Knowledge Engineering* 124 (2019), p. 101738.

[100] Siwei Lai et al. "Recurrent convolutional neural networks for text classification". In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[101] Vasileios Lampos and Nello Cristianini. "Tracking the flu pandemic by monitoring the social web". In: *2010 2nd international workshop on cognitive information processing*. IEEE. 2010, pp. 411–416.

[102] Zhenzhong Lan et al. "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942* (2019).

[103] Paul F Lazarsfeld, Robert K Merton, et al. "Friendship as a social process: A substantive and methodological analysis". In: *Freedom and control in modern society* 18.1 (1954), pp. 18–66.

[104] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. 2014, pp. 1188–1196.

[105] Ji Young Lee and Franck Dernoncourt. "Sequential short-text classification with recurrent and convolutional neural networks". In: *arXiv preprint arXiv:1603.03827* (2016).

[106] Jure Leskovec and Eric Horvitz. *Worldwide buzz: Planetary-scale views on an instant-messaging network*. Tech. rep. Citeseer, 2007.

[107] Ji-chao Li et al. "A link prediction method for heterogeneous networks based on BP neural network". In: *Physica A: Statistical Mechanics and its Applications* 495 (2018), pp. 1–17.

[108] Shancang Li, Li Da Xu, and Shanshan Zhao. "5G Internet of Things: A survey". In: *Journal of Industrial Information Integration* 10 (2018), pp. 1–9.

[109] Taisong Li et al. "Deep dynamic network embedding for link prediction". In: *IEEE Access* 6 (2018), pp. 29219–29230.

[110] Xiaoyi Li et al. "A deep learning approach to link prediction in dynamic networks". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM. 2014, pp. 289–297.

[111] Hai Liang and Fei Shen. "Birds of a schedule flock together: Social networks, peer influence, and digital activity cycles". In: *Computers in Human Behavior* 82 (2018), pp. 167–176.

[112] David Liben-Nowell and Jon Kleinberg. "The link-prediction problem for social networks". In: *Journal of the American society for information science and technology* 58.7 (2007), pp. 1019–1031.

[113] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[114] Dong Liu et al. "The network representation learning algorithm based on semi-supervised random walk". In: *IEEE Access* (2020).

[115] Hanwen Liu et al. "Link prediction in paper citation network to construct paper correlation graph". In: *EURASIP Journal on Wireless Communications and Networking* 2019.1 (2019), pp. 1–12.

[116] Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2013.

[117] Liye Ma, Ramayya Krishnan, and Alan L Montgomery. "Latent homophily or social influence? An empirical analysis of purchase within a social network". In: *Management Science* 61.2 (2015), pp. 454–473.

[118] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[119] Ammara Mahmood and Catarina Sismeiro. "Will they come and will they stay? Online social networks and news consumption on external websites". In: *Journal of Interactive Marketing* 37 (2017), pp. 117–132.

[120] Haggai Maron et al. "Provably powerful graph networks". In: *Advances in neural information processing systems*. 2019, pp. 2156–2167.

[121] Adalbert Mayer and Steven L Puller. "The old boy (and girl) network: Social network formation on university campuses". In: *Journal of public economics* 92.1-2 (2008), pp. 329–347.

[122] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[123] Wenjun Mei et al. "Dynamic social balance and convergent appraisals via homophily and influence mechanisms". In: *Automatica* 110 (2019), p. 108580.

[124] Chetan Harichandra Mendhe et al. "A Scalable Platform to Collect, Store, Visualize, and Analyze Big Data in Real Time". In: *IEEE Transactions on Computational Social Systems* (2020).

[125] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[126] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[127] Zachary Miller, Brian Dickinson, and Wei Hu. "Gender prediction on twitter using stream algorithms with n-gram character features". In: (2012).

[128] David Mimno and Andrew McCallum. "Topic models conditioned on arbitrary features with dirichlet-multinomial regression". In: *arXiv preprint arXiv:1206.3278* (2012).

[129] David Mimno et al. “Optimizing semantic coherence in topic models”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, pp. 262–272.

[130] James Moody. “Race, school integration, and friendship segregation in America”. In: *American journal of Sociology* 107.3 (2001), pp. 679–716.

[131] Martina Morris, Mark S Handcock, and David R Hunter. “Specification of exponential-family random graph models: terms and computational aspects”. In: *Journal of statistical software* 24.4 (2008), p. 1548.

[132] Yi Mou and Kun Xu. “The media inequality: Comparing the initial human-human and human-AI social interactions”. In: *Computers in Human Behavior* 72 (2017), pp. 432–440.

[133] Sourjo Mukherjee and Niek Althuizen. “Brand activism: Does courting controversy help or hurt a brand?” In: *International Journal of Research in Marketing* (2020).

[134] Yohsuke Murase et al. “Structural transition in social networks: The role of homophily”. In: *Scientific reports* 9.1 (2019), pp. 1–8.

[135] Öztürk Nazan and Serkan Ayvaz. “Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis.” In: *Telematics and Informatics* 314 (2018), pp. 136–147.

[136] Mark EJ Newman. “Assortative mixing in networks”. In: *Physical review letters* 89.20 (2002), p. 208701.

[137] Viet-An Nguyen, Jordan L Ying, and Philip Resnik. “Lexical and hierarchical topic regression”. In: *Advances in neural information processing systems*. 2019, pp. 1106–1114.

[138] Brendan O'Connor, Michel Krieger, and David Ahn. "Tweetmotif: Exploratory search and topic summarization for twitter". In: *Fourth International AAAI Conference on Weblogs and Social Media*. 2010.

[139] Saffron O'Neill et al. "Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report". In: *Nature Climate Change* 5.4 (2015), pp. 380–385.

[140] Alper Ozcan and Sule Gunduz Oguducu. "Link prediction in evolving heterogeneous networks using the NARX neural networks". In: *Knowledge and Information Systems* 55.2 (2018), pp. 333–360.

[141] Jiaqi Pan et al. "Twitter Homophily: Network Based Prediction of User's Occupation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2633–2638.

[142] Christos H Papadimitriou et al. "Latent semantic indexing: A probabilistic analysis". In: *Journal of Computer and System Sciences* 61.2 (2000), pp. 217–235.

[143] Jeff Passe, Corey Drake, and Linda Mayger. "Homophily, echo chambers, & selective exposure in social networks: What should civic educators do?" In: *The Journal of Social Studies Research* 42.3 (2018), pp. 261–271.

[144] Krunal Dhiraj Patel et al. "Using Twitter for diabetes community analysis". In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 9.36 (2020), pp. 1–16.

[145] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: vol. 12. Oct. 2011, pp. 2825–2830.

[146] Nicola Perra and Santo Fortunato. "Spectral centrality measures in complex networks". In: *Physical Review E* 78.3 (2008), p. 036107.

[147] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. "Investigating robustness and interpretability of link prediction via adversarial modifications". In: *arXiv preprint arXiv:1905.00563* (2019).

[148] Nastaran Pourebrahim et al. "Trip distribution modeling with Twitter data". In: *Computers, Environment and Urban Systems* 77 (2019), p. 101354.

[149] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. "An analysis of the user occupational class through Twitter content". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 1754–1764.

[150] Dinesh Puranam, Vishal Narayan, and Vrinda Kadiyali. "The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors". In: *Marketing Science* 36.5 (2017), pp. 726–746.

[151] Mohiuddin Qudar and Vijay Mago. "A Survey on Language Models". In: (Sept. 2020). `https://www.researchgate.net/publication/344158120_A_Survey_on_Language_Models`.

[152] Mohiuddin Md Abdul Qudar and Vijay Mago. "TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis". In: *arXiv preprint arXiv:2010.11091* (2020).

[153] Mahmudur Rahman and Mohammad Al Hasan. "Link prediction in dynamic networks using graphlet". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 394–409.

[154] Mahmudur Rahman et al. "Dylink2vec: Effective feature representation for link prediction in dynamic networks". In: *arXiv preprint arXiv:1804.05755* (2018).

[155] Delip Rao et al. "Classifying latent user attributes in twitter". In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. 2010, pp. 37–44.

[156] Mirco Ravanelli et al. "Light gated recurrent units for speech recognition". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.2 (2018), pp. 92–102.

[157] Rami Al-Rfou, Bryan Perozzi, and Dustin Zelle. "Ddgk: Learning graph representations for deep divergence graph kernels". In: *The World Wide Web Conference*. 2019, pp. 37–48.

[158] Garry Robins et al. "An introduction to exponential random graph (p*) models for social networks". In: *Social networks* 29.2 (2007), pp. 173–191.

[159] Kyle Robinson and Vijay Mago. "Birds of prey: identifying lexical irregularities in spam on twitter". In: *Wireless Networks* (2018), pp. 1–8.

[160] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

[161] Adam Sadilek, Henry A Kautz, and Vincent Silenzio. "Modeling Spread of Disease from Social Interactions." In: *ICWSM*. Citeseer. 2012, pp. 322–329.

[162] Adam J Saffer, Aimei Yang, and Maureen Taylor. "Reconsidering power in multi-stakeholder relationship management". In: *Management Communication Quarterly* 32.1 (2018), pp. 121–139.

[163] Mannila Sandhu, Philippe J Giabbanelli, and Vijay K Mago. "From social media to expert reports: The impact of source selection on automatically validating complex conceptual models of obesity". In: *International Conference on Human-Computer Interaction*. Springer. 2019, pp. 434–452.

[164] Abdolreza Shaghaghi, Raj S Bhopal, and Aziz Sheikh. "Approaches to recruiting 'hard-to-reach'populations into research: a review of the literature". In: *Health promotion perspectives* 1.2 (2011), p. 86.

[165] Neel Shah, Darryl Willick, and Vijay Mago. "A framework for social media data analytics using Elasticsearch and Kibana". In: *Wireless networks* (2018), pp. 1–9.

[166] Neel Shah et al. "Assessing Canadians Health Activity and Nutritional Habits Through Social Media". In: *Frontiers in Public Health* 7 (2020), p. 400.

[167] Parag Singla and Matthew Richardson. "Yes, there is a correlation: -from social networks to personal behavior on the web". In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 655–664.

[168] Tom AB Snijders. "Markov chain Monte Carlo estimation of exponential random graph models". In: *Journal of Social Structure* 3.2 (2002), pp. 1–40.

[169] Kihyuk Sohn. "Improved deep metric learning with multi-class n-pair loss objective". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 1857–1865.

[170] Thorvald Sørensen et al. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons". In: (1948).

[171] Mark Steyvers and Tom Griffiths. "Probabilistic topic models". In: *Handbook of latent semantic analysis* 427.7 (2007), pp. 424–440.

[172] Alex Stivala, Garry Robins, and Alessandro Lomi. "Exponential random graph model parameter estimation for very large directed networks". In: *PloS one* 15.1 (2020), e0227804.

[173] Sina Tabakhi, Parham Moradi, and Fardin Akhlaghian. "An unsupervised feature selection algorithm based on ant colony optimization". In: *Engineering Applications of Artificial Intelligence* 32 (2014), pp. 112–123.

[174] Nadine Tamburrini et al. "Twitter users change word usage according to conversation-partner social identity". In: *Social Networks* 40 (2015), pp. 84–89.

[175] Duyu Tang, Bing Qin, and Ting Liu. "Document modeling with gated recurrent neural network for sentiment classification". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1422–1432.

[176] Jian Tang et al. "Line: Large-scale information network embedding". In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 1067–1077.

[177] Jiliang Tang et al. "Exploiting homophily effect for trust prediction". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 53–62.

[178] Joseph Tassone et al. "Utilizing Deep Learning to Identify Drug Use on Twitter Data". In: *arXiv preprint arXiv:2003.11522* (2020).

[179] Jean-Marc Valin and Jan Skoglund. "LPCNet: Improving neural speech synthesis through linear prediction". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 5891–5895.

[180] Tyler J VanderWeele. "Sensitivity analysis for contagion effects in social networks." In: *Sociological Methods & Research* 54.13 (2017), pp. 3058–3070.

[181] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. "Relational Deep Learning: A Deep Latent Variable Model for Link Prediction." In: *AAAI*. 2017, pp. 2688–2694.

[182] Peng Wang et al. "Link prediction in social networks: the state-of-the-art". In: *Science China Information Sciences* 58.1 (2015), pp. 1–38.

[183] Keith Warren et al. "Building the community: Endogenous network formation, homophily and prosocial sorting among therapeutic community residents". In: *Drug and Alcohol Dependence* 207 (2020), p. 107773.

[184] Olga Wichrowska et al. "Learned optimizers that scale and generalize". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3751–3760.

[185] Georg Wiese, Dirk Weissenborn, and Mariana Neves. "Neural Domain Adaptation for Biomedical Question Answering". In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 2017, pp. 281–289.

[186] et al Williams Hywel TP. "Network analysis reveals open forums and echo chambers in social media discussions of climate change." In: *Global environmental change* 32 (2015), pp. 126–138.

[187] Linchuan Xu et al. "Interaction content aware network embedding via co-embedding of nodes and edges". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2018, pp. 183–195.

[188] Sifan Xu and Alvin Zhou. "Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign". In: *Computers in Human Behavior* 102 (2020), pp. 87–96.

[189] Yang Xu et al. "Quantifying segregation in an integrated urban physical-social space". In: *Journal of the Royal Society Interface* 16.160 (2019), p. 20190536.

[190] Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. "Birds of a feather linked together: A discriminative topic model using link-based priors". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 261–266.

[191] Lin Yao et al. "Link prediction based on common-neighbors for dynamic social network". In: *Procedia Computer Science* 83 (2016), pp. 82–89.

[192] Janice Yap and Nicholas Harrigan. "Why does everybody hate me? Balance, status, and homophily: The triumvirate of signed tie formation". In: *Social Networks* 40 (2015), pp. 103–122.

[193] Ussama Yaqub et al. "Analysis of political discourse on twitter in the context of the 2016 US presidential elections". In: *Government Information Quarterly* 34.4 (2017), pp. 613–626.

[194] Matthew D Zeiler. "Adadelta: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701* (2012).

[195] Daojian Zeng et al. "Relation classification via convolutional deep neural network". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 2335–2344.

[196] Chuanting Zhang et al. "Deep learning based link prediction with social pattern and external attribute knowledge in bibliographic networks". In: *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Commu-*

nications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE. 2016, pp. 815–821.

[197] Daokun Zhang et al. "Homophily, structure, and content augmented network representation learning". In: *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE. 2016, pp. 609–618.

[198] Junzhe Zhang and Elias Bareinboim. "Equality of opportunity in classification: A causal approach". In: *Advances in Neural Information Processing Systems*. 2018, pp. 3671–3681.

[199] Long Zhang et al. "Analytical model for predictable contact in intermittently connected cognitive radio ad hoc networks". In: *International Journal of Distributed Sensor Networks* 12.7 (2016), p. 1550147716659426.

[200] Shuai Zhang et al. "Deep learning based recommender system: A survey and new perspectives". In: *ACM Computing Surveys (CSUR)* 52.1 (2019), pp. 1–38.

[201] Xiang Zhang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification". In: *Advances in neural information processing systems*. 2015, pp. 649–657.

[202] Zhongbao Zhang et al. "Efficient incremental dynamic link prediction algorithms in social network". In: *Knowledge-Based Systems* 132 (2017), pp. 226–235.

[203] Chunting Zhou et al. "A C-LSTM neural network for text classification". In: *arXiv preprint arXiv:1511.08630* (2015).

[204] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. "Predicting missing links via local information". In: *The European Physical Journal B* 71.4 (2009), pp. 623–630.

[205] Zhenkun Zhou, Ke Xu, and Jichang Zhao. "Homophily of music listening in online social networks of China". In: *Social Networks* 55 (2018), pp. 160–169.

[206] Jia Zhu et al. "A semi-supervised model for knowledge graph embedding". In: *Data Mining and Knowledge Discovery* 34.1 (2020), pp. 1–20.

[207] Jun Zhu, Amr Ahmed, and Eric P Xing. "MedLDA: maximum margin supervised topic models". In: *Journal of Machine Learning Research* 13.Aug (2012), pp. 2237–2278.

[208] Jun Zhu et al. "Gibbs max-margin topic models with data augmentation". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1073–1110.

[209] Yu-Qian Zhu and Houn-Gee Chen. "Social media and human need satisfaction: Implications for social media marketing". In: *Business horizons* 58.3 (2015), pp. 335–345.

Appendix A

Table of References

Table A.1: The information of all the references selected for the survey

Title	Venue	Citations	Quartile	H-index	Year
Friendship as a social process: A substantive and methodological analysis [103]	Freedom and control in modern society	3069	-	-	1954
Birds of a feather: Homophily in social networks [122]	Annual review of sociology	15216	-	151	2001
Why does everybody hate me? balance,status, and homophily: The triumvirate of signed tie formation [192]	Social Networks	43	Q1	85	2015

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
How social ties transcend class boundaries ? Network variability as tool for exploring occupational homophily [31]	Social Networks	-	Q1	85	2020
Quantifying segregation in an integrated urban physical-social space [189]	Journal of the Royal Society Interface	-	Q1	114	2019
Trustworthy health-related tweets on social media in Saudi Arabia: tweet metadata analysis [6]	Journal of medical Internet research	1	Q1	116	2019
For better or for worse? A systematic review of the evidence on social media use and depression among lesbian, gay, and bisexual minorities [55]	Journal of medical Internet research	6	Q1	116	2019
Social media and human need satisfaction: Implications for social media marketing [209]	Business Horizons	194	Q1	67	2015

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Will they come and will they stay? Online social networks and news consumption on external websites [119]	Business Horizons	194	Q1	67	2015
Political homophily in social relationships: Evidence from online dating behavior [80]	The Journal of Politics	100	-	50	2017
Homophily of music listening in online social networks of China [205]	Social Networks	5	Q1	85	2018
Latent homophily or social influence? An empirical analysis of purchase within a social network [117]	Management Science	85	Q1	221	2016
Twitter Homophily: Network Based Prediction of User's Occupation [141]	Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics	-	-	51	2019
Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter [68]	Journal of public economics	178	Q1	124	2016

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Analyze users' online shopping behavior using interconnected online interest-product network [70]	WCNC	1	-	80	2018
Race, school integration, and friendship segregation in America [130]	American journal of Sociology	1330	Q1	160	2001
Shared contexts, shared background, shared values– Homophily in Finnish parliament members' social networks on Twitter [96]	Telematics & Informatics	3	Q1	52	2019
Building the community: Endogenous network formation, homophily and pro social sorting among therapeutic community residents [183]	Drug and Alcohol Dependence	-	Q1	151	2020
Purity homophily in social networks [47]	Journal of Experimental Psychology: General	83	Q1	138	2016
Structural transition in social networks: The role of homophily [134]	Scientific reports	1	Q1	149	2019

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Information diffusion and opinion change during the gezi park protests: Homophily or social influence? [51]	Database: The Journal of biographical logical Databases and Curation	88	-	65	2016
Trip distribution modeling with Twitter data [148]	Computers, Environment and Urban Systems	2	Q1	74	2019
Good Games, bad host? Using big data to measure public attention and imagery of the Olympic Games [86]	Cities	5	Q1	74	2019
Using Facebook for Qualitative Research: A Brief Primer [56]	Journal of medical Internet research	-	Q1	116	2019
Sensitivity analysis for contagion effects in social networks [180]	Sociological Methods & Researchs	124	Q1	65	2011
Birds of a schedule flock together: Social networks, peer influence, and digital activity cycles [111]	Computers in Human Behavior	3	Q1	137	2018
Social media engagement: What motivates user participation and consumption on YouTube? [90]	Computers in Human Behavior	254	Q1	137	2017

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Social media and web presence for patients and professionals: evolving trends and implications for practice [15]	PM&R	31	-	53	2017
The media inequality: Comparing the initial human-human and human-AI social interactions [132]	Computers in Human Behavior	59	Q1	137	2017
Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis [135]	Telematics and Informatics	80	Q1	52	2018
Network analysis reveals open forums and echo chambers in social media discussions of climate change [186]	Global environmental change	201	Q1	147	2015
The ambivalence of cultural homophily: Field positions, semantic similarities, and social network ties in creative collectives [16]	Poetics	4	Q1	54	2019
A survey on deep learning in medical image analysis	Medical image analysis [113]	2991	Q1	113	2017

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Deep learning in agriculture: A survey [84]	Computers and electronics in agriculture	434	Q1	96	2018
Graph embedding techniques, applications, and performance: A survey [63]	Knowledge-Based Systems	521	Q1	94	2018
Deep learning based recommender system: A survey and new perspectives [200]	ACM Computing Surveys	437	Q1	132	2019
Homophily, structure, and content augmented network representation learning [197]	2016 IEEE 16th international conference on data mining (ICDM)	55	Q1	100	2016
The social structure of political echo chambers: Variation in ideological homophily in online networks [25]	Political Psychology	150	Q1	80	2019
Equality of opportunity in classification: A causal approach [198]	Advances in Neural Information Processing Systems	13	-	54	2018
Yes, there is a correlation: -from social networks to personal behavior on the web [167]	Proceedings of the 17th international conference on World Wide Web	344	-	64	2008

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
The old boy (and girl) network: Social network formation on university campuses [121]	Journal of public economics	443	Q1	123	2008
YouTube vloggers' popularity and influence: The roles of homophily, emotional attachment, and expertise [98]	Journal of Retailing and Consumer Services	-	Q1	65	2020
Dynamic social balance and convergent appraisals via homophily and influence mechanisms [123]	Automatica	1	Q1	239	2019
Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report [139]	Nature Climate Change	151	Q1	136	2015
Twitter users change word usage according to conversation-partner social identity [174]	Social Networkss	47	Q1	85	2015
Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter [99]	Data & Knowledge Engineering	1	Q2	79	2019

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Valence-based homophily on Twitter: Network analysis of emotions and political talk in the 2012 presidential election [72]	New media & society	63	Q1	87	2016
Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign [188]	Computers in Human Behavior	-	Q1	137	2020
Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data [82]	Global Environmental Change	132	Q1	147	2015
Homophily influences ranking of minorities in social networks [85]	Scientific reports	20	Q1	149	2018
Status seeking and perceived similarity: a consideration of homophily in the social servicescape [71]	International Journal of Hospitality Management	29	Q1	93	2017
Emergence of scaling in random networks [12]	Science	36019	Q1	1058	1999

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Spectral centrality measures in complex networks [146]	Physical Review E	148	Q1	190	2008
The influence of cause-related marketing on consumer choice: does one good turn deserve another? [13]	Journal of the academy of marketing Science	1383	Q1	148	2000
Latent semantic indexing: A probabilistic analysis [142]	Journal of Computer and System Sciences	1280	Q2	81	2000
Latent dirichlet allocation [22]	Journal of machine Learning research	31189	Q1	173	2003
Exponential random graph model parameter estimation for very large directed networks [172]	PloS one	4	Q1	268	2020
Measuring the impact of spammers on e-mail and Twitter networks [44]	International Journal of Information Management	10	Q1	91	2019
Friend or frenemy? Experiential homophily and educational track attrition among premedical students [64]	Social Science & Medicine	1	Q1	213	2018

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
A belief-based theory of homophily [89]	Games and Economic Behaviory	5	Q1	84	2019
Tweeting for social justice in# Ferguson: Affective discourse in Twitter hashtags [23]	new media & society	3	Q1	87	2019
An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle [21]	Neurocomputing	7	Q1	110	2019
Semi-supervised classification with graph convolutional networks [95]	arXiv	3196	-	-	2016
An analysis of the user occupational class through Twitter content [149]	Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)	134	-	51	2015
A simple model of homophily in social networks [46]	European Economic Review	66	Q1	116	2016

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Birds of a feather linked together: A discriminative topic model using link-based priors [190]	Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing	5	-	88	2015
Effect of homophily on network formations [93]	Communications in Nonlinear Science and Numerical Simulation	21	Q1	96	2017
MedLDA: maximum margin supervised topic models [207]	Journal of Machine Learning Research	443	Q1	173	2012
Topic models conditioned on arbitrary features with dirichlet-multinomial regression [128]	arXiv	389	-	-	2012
Gibbs max-margin topic models with data augmentation [208]	The Journal of Machine Learning Research	75	Q1	173	2014
Lexical and hierarchical topic regression [137]	Advances in neural information processing systems	61	-	54	2013
Distributed representations of words and phrases and their compositionality [125]	Advances in neural information processing systems	18726	-	54	2013
A density-based method for adaptive LDA model selection [30]	Neurocomputing	276	Q1	110	2009

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
Reading tea leaves: How humans interpret topic models [34]	Advances in neural information processing systems	1668	-	54	2009
Optimizing semantic coherence in topic models [129]	Proceedings of the conference on empirical methods in natural language processing	918	-	88	2011
The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors [150]	Marketing Science	34	Q1	113	2017
Specification of exponential-family random graph models: terms and computational aspects [131]	Journal of statistical software	302	Q1	115	2008
Exponential random graph models for social networks: Theory, methods, and applications [116]	Cambridge University Press	705	-	-	2013
Opening the black box of link formation: Social factors underlying the structure of the web [62]	Social Networks	77	Q1	85	2009

Continued on next page

Table A.1 – *Continued from previous page*

Title	Venue	Citations	Quartile	H-index	Year
An introduction to exponential random graph (p*) models for social networks [158]	Social Networks	1677	Q1	85	2007
Reconsidering power in multi stakeholder relationship management [162]	Management Communication Quarterly	12	Q1	55	2018
Why the Bass model fits without decision variables [17]	Marketing Science	1044	Q1	113	1994
Tweetmotif: Exploratory search and topic summarization for twitter [138]	Fourth International AAAI Conference on Weblogs and Social Media, 2010	411	-	60	2010
A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [170]	Journal of Machine Learning Research	2871	-	-	1948