Social Network Science Approaches for Disease Named Entity Recognition and Extraction

Sarvesh Joshi
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, Mangalore 575025 - India
sarveshjoshi.212it024@nitk.edu.in

Sowmya Kamath S
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, Mangalore 575025 - India
sowmyakamath@nitk.edu.in

***Abstract*—Conventional machine learning approaches adopted for large-scale social media analysis have encountered significant limitations in capturing the underlying dynamics, evolution, and semantic nuances of user posts, hindering comprehensive analysis for tasks related to population health analytics. In this article, the integration of network science-based techniques for node importance/influence analysis, and, Transformer models for Named Entity Recognition are proposed, to facilitate the extraction of structured knowledge from social network posts for population health analytics applications. Standard datasets comprising user account details and postsare considered for the experiments, which are first transformed into graph representations suitable for both structural and behavioral analytics. To evaluate the node importance/influence, different centrality measures were employed and compared. Additionally, a comparative study to assess the impact of varying network sizes by manipulating the number of nodes within the network is conducted. Large-scale mining of disease mentions as a named entity recognition task is also attempted, using neural language models. The proposed approach achieved promising results, outperforming state-of-the-art works by 14.7% in terms of f1-score.**

***Index Terms*—Social Network analysis, Centrality measures, Named Entity Recognition, Population analytics, Network science**

I. Introduction

Social Network Science (SNS) encompasses the investigation of social networks by incorporating principles drawn from network and graph theory. A social network can be visualized as a structure composed of vertices or actors, interconnected through edges representing specific relationships between these entities. These actors can represent a variety of entities, including individuals, organizations, businesses, or nations, with the edges conveying the nature of their connections, through factors like affinity, information exchange, and material transactions. One of the prominent properties, Centrality, is used for understanding networks and their components, from the perspective of a node's significance within a network. Various centrality measures like degree centrality, closeness centrality, and betweenness centrality are commonly employed for this purpose. Degree Centrality measures the number of direct connections a node possesses, pinpointing the node with the highest connectivity within the network. It calculates an importance score based on the node's link count, representing the most straightforward way to gauge node connectivity. Closeness Centrality values are used to evaluate a node's proximity to other nodes in the network, for assessing how efficiently it can help reach other nodes. Betweenness Centrality examines how frequently a node is found on the shortest paths connecting two other nodes. Nodes with high betweenness centrality hold a pivotal role in controlling and facilitating the flow of information or resources within the network, functioning as essential intermediaries.

Natural Language Processing (NLP) techniques are invaluable for extracting insights and trends from text corpora, such as social media posts, user reviews, and news articles, when integrated with SNS. Disease Named Entity Recognition (DNER) aims to identify and extract mentions of diseases or medical conditions from text data, including clinical records, medical literature, social media posts, and news articles. DNER serves several important purposes across various fields, including healthcare, medical research, public health, and data analysis. Its applications include aiding in information retrieval, supporting clinical decision-making, enabling real-time public health surveillance, assisting medical research, facilitating epidemiological studies, contributing to drug discovery, and enhancing public health communication.

The objective of this work is to first consider the tweets as a network graph and find the users that are most influential in a given social network dataset, measured using three different centrality measures - degree, betweenness and closeness centrality. Different network sizes are considered for our analysis, to quantify and assess the changes in influence of a user, based on the changes in social network structure and network dynamics. The rest of this article is structured as follows.Section II discusses some relevant related works with reference to the task being addressed. In Section III, the defined methodology and sub-processes are described in detail. Section IV discusses the implementation, work done and results, followed by conclusion and future work.

II. Related Work

In this section, a discussion of recent research on ML based social network analysis for examining textual data is presented. Virmani et al. [1] explored the concept of social structures in social network data and applied NLP to enhance the accuracy of visualizing the structured information latent in the social

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

network. The natural language text was examined to extract information using a combination of web mining and text mining. Other researchers [2] adopted NLP approaches for modeling social media posts, for tasks like depression detection. Several ML classifiers across multiple domains were trained on data consisting of depressive tweets and Reddit postings, which were modeled using the n-gram character language model (CLM). It was reported that the CLMs for Twitter and Reddit individually had discrepancies in predictions even on similar topics, which they attributed to variations in the underlying structure of the social media platforms.

Network science-based approaches have been utilized for many other tasks. In a study by Namtirtha et al. [3], the authors explored influential individuals in information spreading and introduced a novel approach for their detection. This method combined the sum of neighbor degrees with k-shell decomposition and indexing to find the correlation with the identified influential spreaders. Bail [4] proposed an approach for identifying cultural connections in advocacy-related topics using Autism Spectrum Disorder (ASD) as a case study. The authors found that using cultural bridges in the graph data could yield 2.52 times more relevant user comments when compared to other cases. Shetty and Bhattacharjee [5] studied the growth of COVID-19 pandemic by identifying super-spreading nodes considering the topological structure of the complex network. Their approach considers network nodes; local and global influence, to identify super-spreaders in real-world applications. Kimura et al. [6] emphasized the crucial requirement to take into account node sampling effects on impact indices when attempting to calculate the impact of social media users. Fang et al. [7] developed a novel approach based on Flickr data, wherein, the vertices in the network are users and images, and the relations between them are edges. The performance of photo and friend recommendations was evaluated based on the interactions between the users and images, which resulted in an improvement in accuracy. Antonio et al. [8] utilized PageRank and BiRank algorithms to determine fraud ratings in insurance claim data visualized as graphs. These scores were then employed as features in the supervised learning fraud detection model. They discovered that the model performed better when network and claim-specific variables were combined.

Gao et al. [9] proposed a combination of HDCNN and CRF models for NER from entity tagged data. Their model outperformed the CRF and CNN-CRF models. Li et al. [10] used BERT to provide the word embeddings which is then sent to BiLSTM-CRF layer to generate named entities. They achieved 94% precision which outperformed the other state-of-the-art models. Tong et al. [11] implemented a combination of BiLSTM and CRF modules, and post-processing using Viterbi decoding algorithm for calculating the most optimal state of a word based on previous inputs. They proved that their proposed model performed better than other existing models on NCBI disease corpus.

From the comprehensive review, it is evident that researchers have explored varied approaches for identifying influential nodes within networks, particularly focusing on social media networks and estimating node influence within those networks. Different researchers have focused on the role of node influence, the presence of cultural bridges, the global/local influence of nodes within a network, etc. However, these works overlooked how node influence changes with variations in network size and how the evolving dynamics of the network can introduce new influential nodes and potentially affect previously influential ones. In our work, we aim to focus on this aspect, to model the inherent characteristics and interaction between the nodes, for health domain related tasks.

III. Proposed Methodology

A. Dataset Specifics

The dataset considered for the experimental validation of the proposed approach is a tweet dataset related to disease mentions named SocialDisNER [12], which is provided for mining social media content for disease mentions[1]. The dataset is in Spanish language and contains 5000 tweets in the training, 2500 in validation, and 25000 in the test dataset. These tweets contain disease name mentions among all the text in the tweet, which are in the Spanish language. Table I shows the details of the dataset. The dataset is generated by extracting the tweet ID from the text file name and the corresponding tweet that it contains. Then with the help of Googletrans library for Python [2], the Spanish language tweets are converted to English language. A dataframe is now generated that contains 'tweet id', 'spanish text' and 'english text'.

TABLE I
Dataset Specifics

Parameter	Training	Development
# Tweets	5,000	2,500
# characters	1,253,431	516,768
# tokens	211,555	84,478
Avg. characters per tweet	250.69	206.71
Avg. tokens per tweet	42.31	33.79
# disease mentions	15,173	4,252
# unique disease mentions	4,407	1,413

B. Data Preprocessing

In the data cleaning process, it is crucial to address the presence of redundant characters, URLs, and other irrelevant information commonly found in social media data. To tackle this, the Regex Python library [13], known for its regular expression capabilities, was employed for data preprocessing. Lowercasing of tweets was performed to ensure uniformity and ease of analysis. Removal of URLs, usernames, hashtags, and digits was carried out to eliminate irrelevant elements that could potentially affect the quality of the data. However, when it came to words containing hashtags, complete elimination was avoided as these words could potentially include relevant disease names. The final stage of data preprocessing involved

[1] https://temu.bsc.es/socialdisner/

[2] https://pypi.org/project/googletrans/

the use of the nltk library stopword list, which is specifically designed to eliminate stopwords in the English language. This step aims to remove commonly used words that do not contribute significantly to the overall analysis, thereby refining the dataset for further processing and analysis.

C. Graph Construction

This step involves the transformation of a given Twitter dataset into a network graph representation. This process involves using the tweet IDs in the dataset to acquire the corresponding user IDs. This task is facilitated by making use of the functionalities offered by the tweepy library[3], which enables the retrieval of user details from Twitter. As a result, a JSON object is generated, from which the individual user IDs are extracted. These user IDs are then employed to retrieve the followers linked to each user. The retrieval of follower IDs is accomplished using the "get_follower_ids" function. This process is employed to build a dataframe in which the initial column denotes the source user ID, while the second column contains the user IDs of their followers. Afterward, this assembled dataset is utilized to construct a network graph representation by leveraging the capabilities of the networkx library[4]. This library provides a comprehensive set of tools and functions for both analyzing and visualizing graphs. The resultant network graph offers a visual portrayal of the connections among users within the Twitter dataset, facilitating additional analysis and the extraction of insights.

D. Network Analysis

Next, the generated network is analysed based upon the centrality values of the nodes computed from the constructed graph. Degree centrality determines the most connected node in the group by counting the number of direct linkages to each node. The higher the value the more influential the node becomes. Degree centrality is calculated as per Eq. (1), where, d_v is node v's degree and N represents all nodes in network. Closeness centrality for a particular node can be calculated by calculating the length of all the shortest paths from that node to all other nodes and then averaging it. Closeness centrality is calculated as per Eq. (2), where, $R(v)$ represents the range of node v. Betweenness centrality of a node refers to as the number of times a shortest path passes through that node, over the number of total shortest paths. Betweenness centrality as per Eq. (3), where, $\sigma_{s,t}$ represents the count of shortest paths between s and t. $\sigma_{s,t}\,(v)$ represents the count of shortest paths between s and t that pass through v.

$$Centrality_{degree}\,(v) = \frac{d_v}{|N|-1} \tag{1}$$

$$Centrality_{closeness}\,(v) \;=\; (\frac{|R(v)|}{|N|-1}) * (\frac{|R(v)|}{\sum_{u \in R(v)} d(v,u)}) \tag{2}$$

$$Centrality_{betweenness}(v) = \sum_{s,t \in N} (\frac{\sigma_{s,t}(v)}{\sigma_{s,t}}) \tag{3}$$

To explore the impact of these centrality measures, calculations were conducted for network graphs of various sizes. Specifically, the analysis was performed on graphs with 50, 100, 200, 300, 400, and 500 nodes. This allows for the examination of how centrality values vary with the scale of the network, providing valuable insights into the structural and influential characteristics of different-sized networks. The results of this analysis are presented in Section IV.

E. Named Entity Recognition

The disease-named entity recognition (DNER) is a crucial task in natural language processing that involves identifying and extracting disease-related terms and concepts from the text. It is essential to medical information retrieval and clinical decision support systems. The goal of DNER is to automatically identify and classify diseases mentioned in text into specific categories, such as symptoms, treatments, or diagnoses. In this phase, the disease names are extracted from the given tweet dataset. Every word or group that constantly refers to the same item is considered an entity, each recognized object is put into a specific category. Towards this, different models are adopted for extracting the disease names from the SocialDisNER dataset. The starting and ending of the particular disease names were extracted with the help of regular expressions. The deep learning and transformer models employed for identifying and extracting the required entities for disease NER are ScispaCy, HunFlair, and Spanish BERT models, which are trained on the generated corpus.

1) **ScispaCy model**: ScispaCy [14] is built on top of the spaCy library that is specifically designed for biomedical text analysis to identify entities that correspond to diseases. The complexity of medical literature, including the use of acronyms, technical terms, and intricate sentence structures, can make this task challenging. The 'en_ner_bc5cdr_md' model uses a blend of convolutional and recurrent neural networks and is particularly well-trained to recognize biochemical named entities.
2) **HunFlair model**: Hunflair [15] integrates BiLSTM and CRF models, where, the BiLSTM component captures forward and backward contextual information to generate a coherent and accurate output, while the CRF component models the dependencies between labels for improved resilience. The model also contains gene/protein, chemical, illness, species, and cell line models for precise classification of extractions.
3) **BERT model**: A BERT-based Named Entity Recognition (NER) model is proposed for efficient recognition and categorization of named entities in Spanish language texts. The model is trained on a Spanish dataset of disease names mentioned in tweets using the BIO tagging scheme [16], where B stands for begin, I for inside and O for outside. The proposed method addresses challenges such as ambiguity, word sense disambiguation,

[3] https://docs.tweepy.org/en/stable/
[4] https://networkx.org/documentation/stable/release/release_2.8.8.html

and cross-lingual named entity recognition in Spanish text. Evaluation metrics are used to select the best-performing model.

IV. Experimental Results

Fig. 2 illustrates the interconnectedness of the graph, indicating that followers of a user tend to follow other users as well. However, there are some users who have fewer connections and are located farther from the network's center. These users exhibit lower connectivity within the network. Fig.4 displays the degree distribution of the graph derived from the tweet network. The distribution follows the Pareto distribution, also known as the 80-20 rule. This means that 80% of the nodes have a lower degree and contribute less to the overall degree centrality value. In contrast, 20% of the nodes have a higher degree and significantly impact the degree centrality value. Notably, Fig.4 highlights that nodes with a degree of 1 are more numerous compared to nodes with a higher degree.

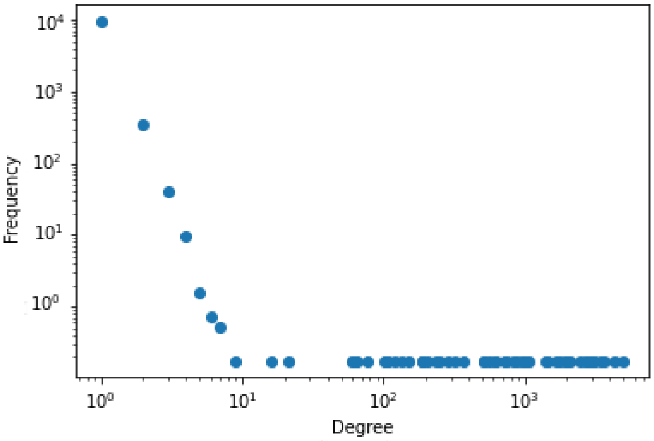

Fig. 1. Degree distribution graph for 50 nodes

We conducted exploratory data analysis (EDA) on the preprocessed textual dataset. Firstly, we created a wordcount dictionary to store the frequency of each word in the dataset. Subsequently, we developed another dictionary that contained the filtered words, which were the words obtained after removing stopwords from the dataset. The word cloud shown in Fig. 2 displays the filtered word dictionary values. This representation of the word cloud was more focused on the medical domain, which is well-suited for our application. In the word cloud, it can be observed that the disease names are presented in large text, while the other words are in smaller text. Fig. 3 represents users who have tweeted disease names more than 10 times. It is evident that the user "Diabetes Foro" has the highest number of tweets with 39 tweets, whereas the remaining users have fewer than 20 tweets. The majority of users appear to have tweeted only once, as in the medical domain, users tend to maintain anonymity. It was also observed that users with many tweets are typically medical institutions or government agencies.

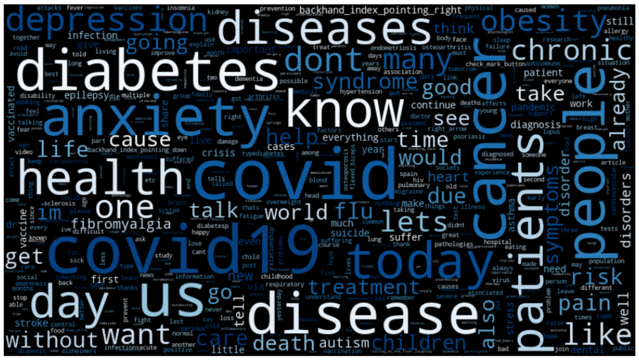

Fig. 2. Word Cloud after stopword removal

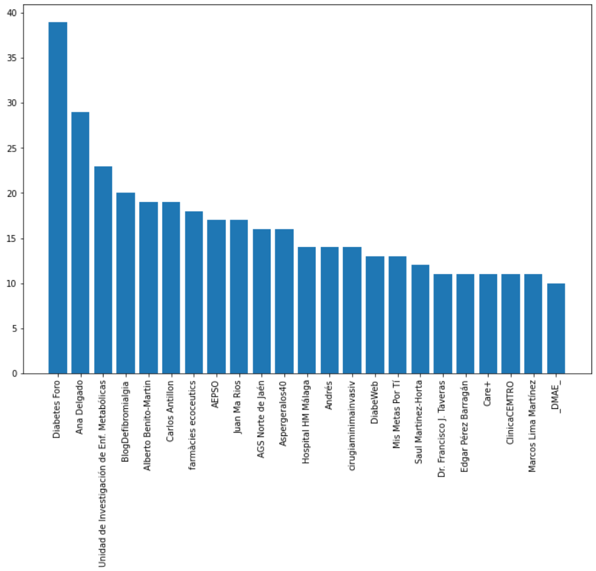

Fig. 3. Disease-mention tweet statistics

Table II provides insights on the characteristics of the graph network when the number of users in the network are increased. Here, it can be clearly observed that as the number of users in the network increase, the number of edges and the number of vertices between them also increase, along with the average degree of the network. The graphs of different size are all sparse graph which is in turn a general characteristic of social network graphs. Another interesting observation is that, while the maximum degree of a node is 5000, most of the users degree is 1 so the average degree is much closer to 1. Also, as the number of users increases the number of connected components also increase, which may be a contributing factor towards the change in the influence of nodes within the network. The average clustering coefficient keeps on increasing as the network size increases, but on the whole, the average clustering coefficient is low which indicates that in general the networks considered here are not scale-free, but more random in nature. Some of the nodes form clusters or tightly interconnected groups, which leads to some users having relatively high influence over their particular cluster within a network.

TABLE II
CHARACTERISTICS OF NETWORK ACCORDING TO NUMBER OF NODES

Number of users	Number of Nodes	Number of Edges	Average Degree	Number of Connected Compo-nents	Average Clustering Coefficient
50 users	58085	60821	2.0942	1	0.0093
100 users	130286	137788	2.1151	2	0.0094
200 users	260040	292134	2.2468	3	0.0197
300 users	357401	410224	2.2955	2	0.0220
400 users	457157	546940	2.3927	4	0.0254
500 users	537419	657863	2.4482	6	0.0280

TABLE III
COMPARISON OF CENTRALITY VALUES FOR VARYING NUMBER OF NODES

Number of users	Degree Centrality	Closeness Centrality	Betweenness Centrality	Type of User
50 users	0.06086	**0.45017**	**0.33203**	PU
	0.06510	0.39218	0.25400	PO
100 users	**0.06510**	**0.39218**	**0.25400**	PO
200 users	0.01612	**0.37835**	**0.11391**	PO
	0.01924	0.37694	0.10742	PO
300 users	**0.01400**	**0.36694**	0.07336	PO
	0.01172	0.36603	**0.08161**	PO
400 users	0.00916	0.35743	**0.05352**	PO
	0.01095	**0.36569**	0.04823	PO
500 users	**0.04684**	**0.34269**	0.03733	GO
	0.03614	0.34184	**0.04181**	PO

For the experimental study, the data was categorized into three different types of users, that is, 'Private User (PU)', 'Private organization (PrO)' and 'Public organizations (PuO)'. Here, the group 'private organizations' indicates the users who are associated with some of the private organizations such as hospitals, newsrooms, journalists, etc., while the group 'public organizations' includes the users that are linked to government organizations like government hospitals, universities, public groups formed to fight a particular disease, etc. Lastly, PUs indicate the individual users present in Twitter that are not linked to any organization.

Table III shows a comparison between the different centrality measures of the network. In a network comprising 50 users, it is evident that one of the most influential user types is the PU, while the majority are PO. Notably, this situation, where a PU is among the most influential users, occurs exclusively in this particular scenario. As the network size expands, the most influential users predominantly belong to the category of private organizations in nearly all cases. Additionally, it's noticeable that private hospitals tend to have a more significant presence on social media compared to government hospitals/organizations (GO). Furthermore, during disease outbreaks, the primary sources of information are newspapers and journalists, categorized as Private organizations. Only in the case of a 500-user network can we observe that one of the most influential users falls under the government user category. In this instance, the influential user is a public help group associated with a specific disease.

There is a significant variation in degree centrality for the top 10 most influential users, as the degree centrality value of a node varies when the size of the graph is altered and when new nodes enter the network. We observe that the value of degree centrality for 100, 200, 300, 400 users keeps on decreasing but the value increases in case of 500 users. This may be due to the fact that the graph becomes more disconnected in the case of 500 users and the average degree of the graph also increases. There is not much variation in closeness centrality as the number of users of the network change, but as new nodes get added to the network such nodes become more influential as their closeness centrality value increases. Betweenness centrality decreases as the number of nodes increases. One of the reasons behind this is that as the number of nodes in the network increases the network becomes more disconnected, so many important links within the network might be broken as new nodes enter the network thus altering the shortest paths of the network.

Next, three different models, SciSpacy, Hunflair and BERT are adopted for the tasks of Disease Named Entity Recognition (DNER). The deep learning models were used on the Spanish tweet dataset which is converted to English using 'Googletrans' library. Regular expressions are used to identify the beginning and ending of disease names. For the experimental evaluation, various standard metrics like precision, recall, and f1-score were used. Precision is measured as the ratio of predictions made by the model to the actual number of predictions that are actually correct. Recall measures how many positive predictions were made over all the positive instances in the data. These are computed by considering the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values. F1-score combines both precision and recall, a higher precision and recall value results in higher F1-score. Equations (4), (5) and (6) show the formulae used to compute precision, recall and F1-score values.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

To observe performance, 25 distinct BERT models are trained by altering the number of epochs and learning rates. The models are formed by combining five different epochs (ranging from 5 to 9) and five different learning rates (0.00005, 0.000005, 0.0000005, 0.00000005, 0.000000005). The hyperparameters of the BERT model were observed for 5, 6, 7, 8, and 9 epochs, and also for various learning rates. Table IV shows the results of these experiments. The model name is indicated in the form of Ep5lr1, where Ep stands for epochs, the number next to it represents the number of epochs, $lr1$ represents learning rate of 0.00005. Similarly $lr2$ = 0.000005, $lr3$ = 0.0000005, lr4 = 0.00000005 and lr5 = 0.000000005.

From the results, it can be discerned that the evaluation metrics exhibit an improvement when different BERT models are employed, which have been trained using our specific dataset. The performance is affected significantly when the learning rate is reduced. In contrast, the model did not train well when the learning rate was increased. Table IV indicates that the combination of using 9 epochs and 0.00005 as the learning rate produced the overall best value for all metrics. The highest precision value was obtained for the model trained for 9 epochs and the learning rate of 0.00005. The best F1-score value was achieved by the same combination of 9 epochs and 0.00005 learning rate, while the highest recall value was obtained by the model trained for 7 epochs and 0.00005 learning rate.

TABLE IV
EVALUATION METRICS FOR DIFFERENT BERT MODELS

Model	Epochs	Learning Rate	Precision	Recall	F1-Score
Ep5Lr1	5	0.00005	0.909	**0.917**	0.913
Ep5Lr2	5	0.000005	0.9	0.911	0.906
Ep5Lr3	5	0.0000005	0.854	0.857	0.856
Ep5Lr4	5	0.00000005	0.805	0.713	0.756
Ep5Lr5	5	0.000000005	0.825	0.651	0.728
Ep6Lr1	6	0.00005	0.919	0.906	0.913
Ep6Lr2	6	0.000005	0.906	0.909	0.907
Ep6Lr3	6	0.0000005	0.866	0.867	0.867
Ep6Lr4	6	0.00000005	0.803	0.719	0.759
Ep6Lr5	6	0.000000005	0.824	0.652	0.728
Ep7Lr1	7	0.00005	0.914	**0.917**	0.915
Ep7Lr2	7	0.000005	0.905	0.909	0.907
Ep7Lr3	7	0.0000005	0.869	0.873	0.871
Ep7Lr4	7	0.00000005	0.806	0.728	0.765
Ep7Lr5	7	0.000000005	0.825	0.653	0.729
Ep8Lr1	8	0.00005	0.915	0.912	0.913
Ep8Lr2	8	0.000005	0.902	0.91	0.906
Ep8Lr3	8	0.0000005	0.875	0.877	0.876
Ep8Lr4	8	0.00000005	0.806	0.736	0.77
Ep8Lr5	8	0.000000005	0.825	0.653	0.729
Ep9Lr1	9	0.00005	**0.924**	0.911	**0.917**
Ep9Lr2	9	0.000005	0.909	0.91	0.909
Ep9Lr3	9	0.0000005	0.879	0.88	0.879
Ep9Lr4	9	0.00000005	0.805	0.744	0.773
Ep9Lr5	9	0.000000005	0.823	0.655	0.729

V. CONCLUSION AND FUTURE WORK

In this paper, a publicly available social media dataset containing disease mentions was considered for the task of network science based analysis of network dynamics and node influence. The dataset was preprocessed and modeled to generate its network representations. Using the tweet ids and their followers, the users' network centrality measures were calculated and a comparative study based on the different sizes of networks was performed. The study revealed that as the size of the network increases some of the newer nodes become more influential than the previously influential nodes while some nodes are influential in all the different sizes of the network. We also observed that, as the size of the network increases the network becomes more disconnected. The nodes which were influential in a smaller community of nodes, lose influence when the size of the network is increased, due to this disconnection. Other nodes whose influence does not decrease are more influential in larger community. So, in case of disconnected graph the influence of a node depends on the size of the community it is present in and if it is linked to any other community by any bridge or not. If it is linked by a bridge then its importance increases. Further, a publicly available dataset containing mentions of diseases was utilized for predicting disease name mentions on social media posts. Since the dataset was originally in Spanish, we performed a language conversion to English to make it compatible with our task. Subsequently, pretrained deep learning models were employed for NER task. BERT was trained exclusively on Spanish data and its hyperparameters were fine-tuned, obtaining a precision of 0.924, the best-in-class performance. We intend to explore other transformer models for varied multilingual DNER tasks.

REFERENCES

1 Virmani, C., Pillai, A., and Juneja, D., "Extracting information from social network using nlp," *International Journal of Computational Intelligence Research*, vol. 13, no. 4, pp. 621–630, 2017.

2 Marnauzs, S. and Kalita, J., "A domain independent social media depression detection model," 2019.

3 Namtirtha, A., Dutta, A., Dutta, B., Sundararajan, A., and Simmhan, Y., "Best influential spreaders identification using network global structural properties," *Scientific Reports*, vol. 11, no. 1, Jan 2021.

4 Bail, C. A., "Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media," *National Academy of Sciences*, vol. 113, no. 42, 2016.

5 Shetty, R. D., Bhattacharjee, S., and Dutta, A., "Gsi: An influential node detection approach in heterogeneous network using covid-19 as use case," *IEEE Transactions on Computational Social Systems*, pp. 1–15, 2022.

6 Kimura, K. and Tsugawa, S., "Estimating influence of social media users from sampled social networks," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1302–1308.

7 Fang, Q., Sang, J., Xu, C., and Rui, Y., "Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 796–812, 2014.

8 Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., and Reynkens, T., "Social network analytics for supervised fraud detection in insurance," *Risk Analysis*, vol. 42, no. 8, 2022.

9 Gao, M., Wei, H., Chen, F., Qu, W., and Lu, M., "Hdcnn-crf for biomedical text named entity recognition," in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, 2019.

10 Li, W.-Y., Song, W.-A., Jia, X.-H. *et al.*, "Drug specification named entity recognition base on bilstm-crf model," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, 2019, pp. 429–433.

11 Tong, F., Luo, Z., and Zhao, D., "A deep network based integrated model for disease named entity recognition," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

12 Gasco Sánchez, L., Estrada Zavala, D. *et al.*, "The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora," in *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, 2022.

13 Lundh, F. and Kuchling, A. M., "Re - regular expression operations," Oct 2012. [Online]. Available: https://docs.python.org/3/library/re.html

14 Neumann, M., King, D., Beltagy, I., and Ammar, W., "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Aug. 2019.

15 Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., and Akbik, A., "Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition," *Bioinformatics*, vol. 37, no. 17, 2021.

16 Ramshaw, L. A. and Marcus, M. P., "Text chunking using transformation-based learning," 1995.