Extracting Features from Online Forums to Meet Social Needs of Breast Cancer Patients

Maitreyi Mokashi, Enming Zhang, Josette Jones, Sunandan Chakraborty
[mmokashi,enzhang,jofjones,sunchak]@iu.edu
School of Informatics and Computing
Indiana University Purdue University Indianapolis
Indianapolis, IN

ABSTRACT

Breast cancer patients go through many ordeals when they undergo treatments. Many of these issues are personal, social, or professional. As many of them are not directly medical in nature, these issues are not discussed with their healthcare providers and hence, not included in their treatment plan. However, these issues are vital for the patients' complete recovery. We present a novel approach that acts as the first step in including such personal and social issues resulting from breast cancer treatment into a patient's treatment plan. There are numerous online forums where patients share their experiences and post questions about their treatments and subsequent side effects. We collected data from one such forum called "Online Breast Cancer Forum". On this forum, users (patients) have created threads across many related topics and shared their experiences and questions. We use these message threads to identify critical issues faced by the patient and how they are related to their treatment. We convert the forum data into a bipartite network and turn the network nodes into a high-dimensional feature space. In this feature space, we perform community detection to unearth latent connections between patients and topics. We claim that these latent connections, along with the known ones, will help to create a new knowledge base that will eventually help physicians to estimate non-medical issues for a prescribed treatment. This new knowledge will help the physicians plan a more adaptive and personalized treatment and be better prepared by anticipating potential problems beforehand. We evaluated our method on two baseline methods and show that our method outperforms the baseline methods by 25% on a manually labeled reference dataset.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; • **Human-centered computing** → **Collaborative and social computing**; • **Information systems** → *Extraction, transformation and loading*; **Wrappers (data mining)**; Web searching and information discovery.

KEYWORDS

Health informatics, Text mining, Network embedding, Social computing

ACM Reference Format:
Maitreyi Mokashi, Enming Zhang, Josette Jones, Sunandan Chakraborty. 2020. Extracting Features from Online Forums to Meet Social Needs of Breast Cancer Patients. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '20), June 15–17, 2020, , Ecuador*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3378393.3403652

1 INTRODUCTION

"The ONE thing I've wanted my entire life - a long lasting successful relationship - is the one thing I don't ever get to have. Could I meet a guy someday that might ask me out? Sure. What's going to happen when he hears my diagnosis? He'll run for the hills."

This message was posted on an Online Breast Cancer Forum (OBCF) - Breastcancer.org. This post captures a strong message about a breast cancer patient's well-being which is dependent on many factors, many of which are beyond the scope of discussions that take place in the confines of a clinic. Cancer patients undergo numerous ordeals − effects of the disease and the side effects of its treatments. Many of these issues are not medical but personal or social in nature. As a result, patients do not think these issues as relevant and refrain from discussing them with their health-care providers. Moreover, patients also refrain from discussing such issues due to the stigma associated with it [29].

Acknowledging such issues are strongly bound to a patient's complete recovery. Little is known on how people integrate disease management, especially chronic diseases, into their daily lives. However on the other hand, patients are more comfortable sharing such information on online health forums, social media, and similar platforms to seek solace or advice from fellow patients who are undergoing similar ordeals. The rapid growth of Web 2.0 has made social media a significant platform for health surveillance and social intelligence. Using these platforms, patients form interactive networks by posting and replying to messages, providing reviews and attending discussion boards[16]. In such platforms, patients freely discuss their experience(s) of "I'm not okay" and what they did to "feel okay now". This level of personal information is vital to a patient's path to complete recovery, both physically and mentally.

In this paper, we have introduced a novel approach that can connect personal and social issues associated with a disease, specifically focusing on breast cancer. Treatments and drugs can have adverse effects on a breast cancer patient's daily life. In other words, provide a platform for health care providers to understand and acknowledge the "personal and social issues" a patient might encounter at the

COMPASS '20, June 15–17, 2020, , Ecuador

ACM ISBN 978-1-4503-7129-2/20/06…$15.00
https://doi.org/10.1145/3378393.3403652

current or future state of the disease. Additionally, the physician may get an estimate of what *issues* the patient might face due to the new treatment plan. Our hypothesis is that advance knowledge of such issues will help the physicians provide additional recommendations or referrals for the patients and provide them with a more holistic treatment. We present an approach that will mine this *additional information* from online forums dedicated for breast cancer patients and survivors and represent them as latent features describing a patient. The objective of these features is to succinctly summarize the state of the patient and what social/personal issues she is facing due to this current state.

We collected data from an online health forum dedicated for patients diagnosed or survived breast cancer [1]**(OBCF)**. Registered users can post questions, answer to others' queries and create a supportive community of breast cancer patients and survivors. The questions posted on this forum are from a variety of topics, including but not limited to tests and treatments, diagnosis, immunotherapy as well as the personal matters that the patients face on a daily basis.

The main objective of the proposed approach is to extract information about the key problems patients face and identify the hidden links between these *issues* and the disease factors (e.g. treatment, symptoms). However, this information can come in isolated bits and little fragments shared by individual patients which needs to be aggregated to get the bigger picture. We use the interaction of patients within these forums to find out the commonalities between them and extract the *hidden* links between the state of the disease and all *other* issues. The final outcome of our approach is a *representational framework* for patients. These features will be high-dimensional vector representation of the "state" of the patient, which includes current stage of the disease, treatments, diagnosis and other issues faced by the patient. We learn this patient coding from the interaction of patients in the forum based on an assumption that a patient's interaction is representative of their own issues and experiences. We learn the latent features and use them to represent the connections between the patient and topics, as well as other *similar* patients.

To model the similarity and eventually extract the features, we convert the forum data into a bipartite network. We call this network – *patient-topic network*. In this network, we represent the patients and the forum topics as nodes and the interactions within them as edges. A patient and a topic are connected by an edge, if the patient has participated (posted/replied) in that topic. In other words, an edge represents a relationship between a patient and a topic. Using this notion of similarity across thousands of patients and topics in the forum data, we learn a representational framework that will encode different features of a patient with respect to her diagnosis, treatment and the different (non-medical) issues they are facing resulting from the disease and the treatments. We use node embedding method specially designed for bipartite networks to obtain the embedding vectors. Finally, we perform community detection on the embedded feature space to find clusters of similar patients and topics and identify hidden relationships through these clusters.

We evaluated our method involving the heterogeneous patient-topic network with two baseline models, using a purely text-based approach and a user (patient) network. We measured the coherence of the clusters using normalized pointwise mutual information (NPMI) score [22]. Our model outperformed the two baseline models with an NMPI score of 0.481 compared to 0.237 and 0.294 for the other models. We also measured the similarity of our clusters on a manually created reference dataset [17], which showed that our precision-recall score in identifying correct clusters is around 25% better than the baseline models.

2 RELATED WORK

2.1 Breast Cancer and Social Media

Online forums and social media allows patients to search for health information online and interact with people with similar conditions. During the past decade, many health social media websites were created to facilitate information exchange among patients. Some websites like Everyday Health [2] and PatientsLikeMe [3] cover general health problems and provide information on many aspects of health. Others focus on a particular group of people with similar conditions, such as diatribe's focus on diabetic patients or Disabilities-R-Us' [4] focus on people with disabilities. Many past researches have studied the effect of these online activities and how they support they can provide on patient care and welfare. Previous observational studies have advanced our understanding of how social media helps patients with advice, guidance, and support with their chronic diseases. Greene et al [11] used qualitative methods to evaluate the content of Facebook groups dedicated to diabetes management. This study concluded that a "safe" place to discuss extra-clinical issues helped the patients in general. Apart from these qualitative studies, other observational studies have used data mining and machine learning in their analysis. For example, Park and Ryu [30] applied Natural Language Processing (NLP) methods to extensive online forum text data to understand key problem areas of patients who have fibromyalgia. A similar approach has been used to address a variety of clinical problems, such as public sentiments towards vaccination [18], adverse drug effects [15], influenza epidemic using Twitter data [2], and e-cigarette usage [41].

In Chawla et al [5], emphasize on "Data-driven and networks-driven thinking and methods can play a critical role in the emergence of personalized healthcare." They use a patient-centric model (CARE) that creates a personalized disease risk profile, as well as a disease management plan and wellness plan for an individual. Machine learning and text mining has been extensively used for breast cancer research. Many studies have been done that use clinical data to better diagnosis and treatment of breast cancer [3, 10], early detection of BC [7, 21, 33, 36] as well as for various medical factors such as drugs and treatments [9, 38]. In parallel, numerous studies have focused on the effects of social media on breast cancer research and patients. Modave et al [28] performed sentiment analysis on tweets discussing breast cancer and demonstrated that social media can improve the perceptions of the disease on the general population. Zhang et al [43] used CNNs to extract longitudinal

[1] www.breastcancer.org

[2] https://www.everydayhealth.com

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

[4] https://www.disabilities-r-us.com

information to understand the key topics discussed on online breast cancer forums. There are many factors that determine the proper survival of breast cancer patients, this included access to treatment, financial constraints and many more personal factors. The study by Sheng et al. [37] focused on the long-term effects of these factors on the quality of life of breast cancer patients and survivors.

Trans-disciplinary research has been conducted to create frameworks that take into account social determinants of cancer to observe the environmental, social and behavioral factors of the cancer patients. Hiatt et al [13] designed a framework to conceptualize how social determinants interact with other factors in the etiology of cancer and to capture changes over time. Cancer studies include a complete spectrum of scientific endeavor from genes to society and hence studies may provide pathways for understanding the complex and multilevel causal mechanisms needed to create cancer control and interventions society. Carter et al [4] conduct a qualitative analysis of thirty-two cancer control policy documents, critiquing them based on their likely impact on social determinants and created a matrix and set of questions to guide the development and assessment of health policy.

2.2 Mining on Network Data

Extracting information from online health forums and other instances of user-generated data comes with many challenges. They are noisy, inconsistent and do not use proper structures and formats that usually makes it easy for information extraction. As a result, several data mining techniques have been devised for such data mining tasks. Yang et al. (2012)[39] compare their study with Rossetti et al. [34] , where the later proposed multidimensional versions of the Common Neighbors and Adamic/Adar, and derived predictors that aimed at capturing the multidimensional and edge level temporal information, while the prior gathered nodal historical data to capture the preference of topological features when two nodes are associated by new link; while they are interested in edge level communication data. Network analysis has become a important part of research and organizations as data for example, from social media have a complex structure which form a network via "links". These links help understand the underlying connections and the invisible relations [39]. Grover et al.(2016)[12] introduced a concept for link prediction using Node2Vec which is based on the Word2Vec model by Mikolov et al. 2013 and DeepWalk by Perozzi et al. (2014) [32].

[14, 20, 23, 31] conducted their research by building over the original node2vec model to meet the requirements of the respective study. Li et al. (2017) [23] created a modified version of node2vec by introducing TDL2vec which considers time factor while generating the links during word2vec model. Similarly, Peng et al. [31] created a model to predict Parkinson's disease by creating N2A-SVM algorithm which includes a autoencoder for dimensionality reduction of the node2vec model and Support Vector Machine for the prediction analysis. We propose to use the Node2Vec model and include textual features from the posts made by the users.

3 PROBLEM DEFINITION

The main goal of this paper is to extract features to represent patients. These patient features will form a description of the patients in terms of treatment, symptoms, side effects and other *issues* faced by them. We discover these features from the information shared by them as well as their interaction with different threads in the online forums. We assume that a patient has a direct relationship with a topic in which they have participated, where participation is defined as either posting a new message or replying to an existing one. We formulate the problem with three variables – (1) patients (V), (2) topics[5] (T) and (3) the features (θ). Here, V and T can be observed from data, whereas θ is latent and will be learned.

Our objective is to design a mapping function that encodes patients, their messages and the their interactions into a common feature space. We aim to *bind* patients and the text content together, i.e. *similar* patients with *similar* conditions (or topics) will be placed nearby regions in the high dimension feature space. This similarity is defined by the interaction of the patients in the forums – similar patients will tend to participate on same or similar topics.

We convert the forum data into a bipartite graph. In this graph, we represent the data as $G = (V, T, E)$, where V represents the set of patients and T the set of topics extracted from the forum data. A patient v_i and a topic t_j will be connected by an edge $e_{ij} \in |V| \times |T|$, if the patient v_i have participated (posted/replied) in the topic t_j. As, G is a bi-partite network, there will be edges between patients and topics but no edge between topic or patient pairs. The relationship between two patients or topics is implicit and identifying those implicit relationships is our main goal.

We learn the features through the interaction between patients and topics and also using patient-patient relationships based on participation in common topics. Our goal is to design a mapping function ψ, such that $\psi : V \cup T \rightarrow R^K$, for all patients i and topics j. Here, K represents the dimension of the feature space and it is pre-determined. We used an adaptive node embedding method to formulate this mapping.

The mapping function ψ will produce K-dimensional vectors for each patient and topic and produce an embedding of these two variables. As a result, we will be able to compute the distance between two patients, two topics as well as the distance between a patient and a topic. This way, we are able to discover *implicit* relationships within the network variables that are not directly depicted by the underlying network. While clustering the patients, we will know commonalities between patients and connect issues that are observed under similar conditions but was not known before. At the same time, we will be able to directly connect patients with *nearby* topics and discover potential issues a patient might face given the current state the patient is in. Figure 1 shows an overview of the entire process.

Our approach of extracting the latent features by converting the data into a bipartite network has an advantage of incorporating the patients into the model. Information shared by a single patient about a specific topic t_a is unlikely to contain all possible information about t_a. However, aggregating different patients' experience about t_a from the entire data can provide a more holistic overview of t_a. Thus, including patients as intermediate connecting points is more likely to link many latent connections between the topics. Thus, our claim is that this network-driven method will lead to richer

[5] Here, topics refer to the forum threads as termed by the designer of the Breastcancer.org creators. These topics have no relations with LDA or other statistical topic models

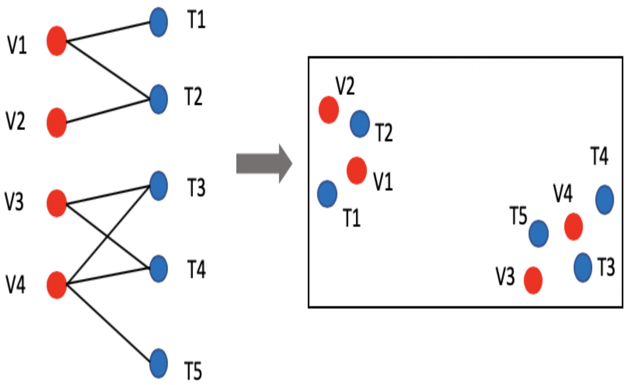

Figure 1: An example patient-topic network and the corresponding embedding. The left side of the figure shows the bipartite network, where red nodes denote patients and topics as blue nodes. As depicted by the network, V1 have direct connection with T1 and T2 and in the feature space V1 is close to both the topics. Whereas, V2 is closer to T2 but further away from T1. On the other hand, T3 and T4 are apart by a greater distance because there no connected link between V1 and V2.

feature extraction compared to a purely text-based approach, where the text-only model will not include the patients as intermediate variables.

4 DATASET

The Online Breast Cancer Forum (OBCF)[6] community provides a platform for the patients and their friends and family to share their experience and post questions etc. This platform hosts multiple forums which are generally specific to one subject [17] and users create new threads to post their questions or opinions related to a specific topic. Figure 2 represents the hierarchical structure of the messages in OBCF.

The levels in this structure (Figure 2) is explained below. The dataset consists of four levels excluding the users. Each level is sub-divided in to various branches based on their context (Table 1, Table 2).

- *Level I* - **Categories**: The OBCF identifies each post as one of 9 different sections for surface-level categorization. This gives the users a rough idea regarding the discussion being done in the subsequent forums and topics (refer Table 1,2).
  - Tests, Treatments and Side Effects
  - Day-to-Day Matters
  - Not Diagnosed but concerned
  - Advocacy and Fund-Raising
  - Community Connections
  - Welcome to Breastcancer.org
  - Site News and Announcements
  - Connecting With Others Who Have a Similar Diagnosis
  - Moving On & Finding Inspiration After Breast Cancer

Table 1: Description of the OBCF dataset

Name	Total
Categories	9
Forums	79
Topics	140000
Replies	4.4 million
Users	94000

Table 2: Analyzing the trend of posts in each level of OBCF dataset.

Posts per type	Forums	Topics	Users
Max	616598	56091	48986
Min	11	1	1
Mean	56304	31	47

- *Level II* - **Forums**: As each 'Category' gives us a surface view of OBCF, 'Forums' sections them further into selective discussions. For example, discussion regarding mastectomy and lumpectomy will most probably be addressed in the 'Breast Reconstruction' forum. We analysed data from 79 forums. Out of approximately 4.4 million posts, one forum *"Day-to-Day Matters"* has 616,598 which is the maximum number of post in one forum. On the other hand, which also indicates that patients see OBCF as a safe place to indulge others with their daily personal issues as well. The average number of posts per forum is 56,300 posts.
- *Level III* - **Topics**: When a user has a new subject to discuss or something new to share they post an independent post which directly transmits as a *new topic*. In OBCF, there are 140,000 topics spread across the 79 forums with a maximum of 56,000 replies for one individual topic and a mean of 30 replies. There a few topics with no replies. For example a topic can be: "chemo after a mastectomy" or "Size of tumor by MRI vs Reality" to name a few.
- *Level IV* - **Replies**: All the replies to topics fall under this section. One user has made over 48,000 posts with a mean number of post per user being 47. There is no sub-branching for the replies and all the replies are stacked under this topic (with a respective ID).

In our analysis, we use the entire message thread under a topic to build the bipartite network (G) and represent the merged topic and the stack of replies as one topic (t_j).

Jones et al. used data from **OBCF** to determine the feasibility of acquiring and modeling the topics of this online breast cancer forum [17]. Using qualitative analysis of the QCA obtained topic models and statistical analysis, the obtained topics were placed into 4 distinct clusters and were clinically asserted as significant. A machine learning regression implementation was performed to find highly significant topics. Zhang et al. manually annotated 736 randomly selected posts from **OBCF** and created "Patient-centered Thesaurus of Chronic Survival (PACToCS)", which are then mapped with medical controlled vocabulary - NCI Metathesaurus [42]. The

[6] BreastCancer.org

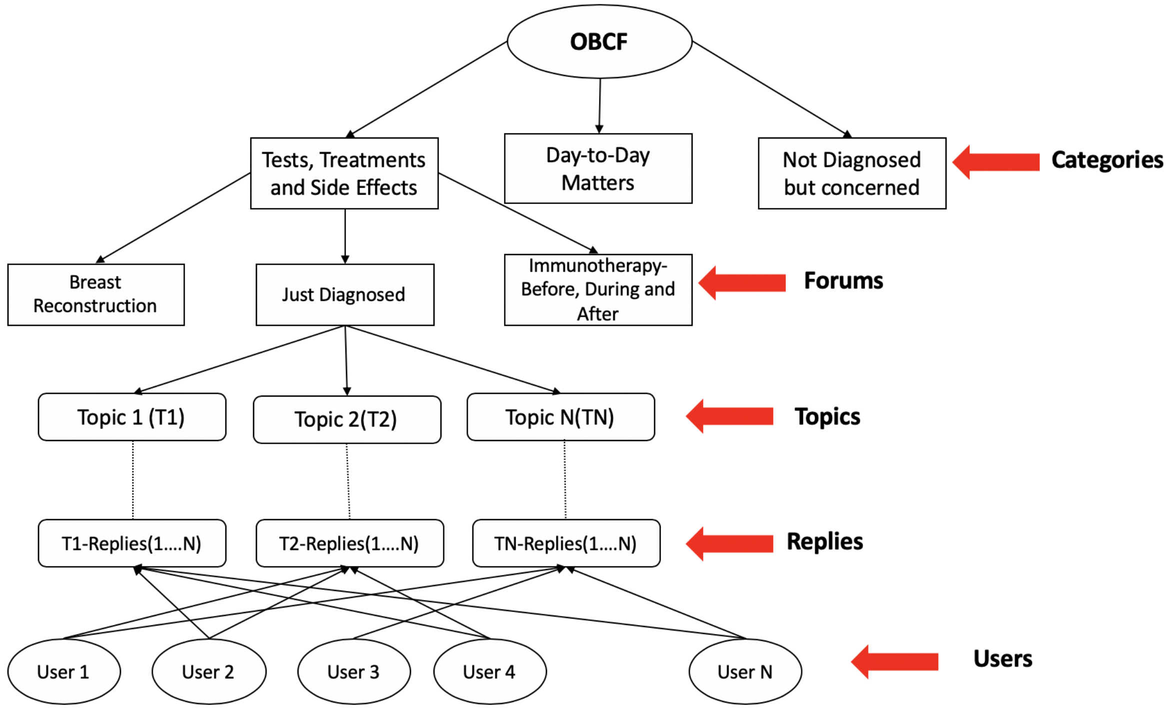

Figure 2: This figure represents the hierarchical structure of the data obtained from OBCF (Breastcancer.org.)

authors identified 30 topics and 27 out of 323 full code terms from PACToCS matched with the full term of the NCI - Metathesaurus. They obtained a precision of 85% upon classification by multiple ML models.

5 METHODOLOGY

5.1 Overview

Our goal is to represent breast cancer patients' state using latent features by analyzing messages posted on an online forum. We represent the forum data as a bipartite network that represents the explicit relationship between a patients and shared topics as well as to capture the implicit relationship between the patients and topics that are not directly depicted in the network. Through our method we wish to quantify the patients' experience across different stages of the disease and encode that information into a high-dimensional vector that can embed a variety of information, including diagnosis, treatment, side effects of the treatment as well as mental/social issues. This representational model will also help to identify similar patients and plan a personalized treatment plan for new incoming patients.

5.2 Patient-Topic Network

Each breast cancer patient experiences a unique set of challenges. Looking at an individual patient in isolation will make one part of the story visible. For example, a patient v_i shares an information related to a treatment option b for the disease state a, we aim to identify all issues about b across the dataset. In our dataset, the information related to a single post about a and b are represented as one topic t_j. In addition, our goal is to capture other issues shared by patients who are *similar* to v_i. This will provide a more holistic view of the issues faced by a patient who are in the same state as a and who are receiving the treatment b. Some of these relationships are explicit and many of them are implicit and needs special action to identify them. We learn the latent features from how the patients interact on the forums. Thus, we convert the data into a patient-topic bipartite network.

We convert the forum data into a network G and call this the patient-topic network. We represent the forum topics as T. A topic $t_j \in T$ represents a new post and the stack of replies. A topic t_j will have T_j^N posts, including the original post and T_j^{N-1} replies. $V = \{v_1, v_2, v_3, v_4, ..., v_M\}$ represents the set of M patients, thus, there are $M + N$ nodes in the patient-topic network G. If a patient v_i has posted in t_j, i.e. has posted $t_j^k \in t_j$, where $1 \leq k \leq t_j^N$, there will be an edge between v_i and t_j. Figure 1 shows an example patient network.

5.3 Network Embedding

Our goal is to preserve the network structure and find a representation of the graph as a real-valued vector. The resulting vector representation will place neighboring nodes (i.e. patients) closer to each other while placing others far apart in the high dimension vector space. Unlike homogeneous networks, in a bipartite network the same type of node are not connected by an edge. In our case, two topic nodes or two patient nodes are not directly connected. However, that does not imply that those nodes are not related. This poses an additional constraint on the learning objective of the node embedding in our case and standard node embedding methods [12, 32] are not directly applicable. We use the specialized embedding method designed for bipartite network – BiNe [8]. BiNE is designed to utilize the heterogeneous nature of the network and can measure proximity even when the nodes are not connected by an edge, i.e. when the two nodes are of same type. Using the BiNE method we are able to model both explicit and implicit relations within the patient-topic network. In our case, explicit relationships are depicted by edges connecting a pair of patient, topic nodes. On the other hand, two patient nodes who are connected by an intermediate topic or two topic nodes connected similarly are examples of implicit relationships (i.e. not connected by an edge but still they are proximate).

Similar to BiNE model, we model the explicit relations as the joint probability of two nodes is defined as. If v_i is the i^{th} patient node and t_j is the j^{th}

$$P(t_j, v_i) = \frac{\alpha_{i,j}}{\sum_{e_{ij} \in E} \alpha_{ij}}$$

where, α_{ij} is the edge weights and it represents the frequency of posts of v_i on topic t_j. The objective is to minimize the KL-divergence of the actual measure of the explicit relationship (P) and the expected value ($\hat{P}$) computed from the vector representation t_j and v_i, represented as ω_j and ω_i respectively. $\hat{P}$ is computed as,

$$\hat{P}(v_i, t_j) = \frac{1}{1 + e^{-\omega_j^T \omega_i}}$$

Implicit relationships along with explicit ones are useful to extract features from bipartite networks. In our case, the implicit relationships are key in identifying hidden connections between patients and topics, even when the patient has no direct connection with the topic. No direct connection between a patient (v_i) and a topic (t_j) means that this particular patient has not shared or participated on that topic. However, they still can have a relationship with that topic t_j through other *similar* patients. This is represented in the network as a *short walk* from v_i to t_j via other patients and topics. To model the implicit relationships, we use the random walk method in Node2Vec [12]. Node2Vec uses a mix of Breadth-First Search (BFS) and Depth-First Search (DFS). BFS uses importance of local neighbors or a micro-view, whereas DFS helps to obtain a more spread out connection with a macro-view [12]. The node2vec model uses characteristics from both of these classic search models (Figure 3).

Node2vec model depends on 4 main hyperparameters:

- Number of walks: Number of random walks to be generated from each node in the graph
- Walk length: How many nodes are in each random walk
- P: Return hyperparameter
- Q: Inout hyperparameter

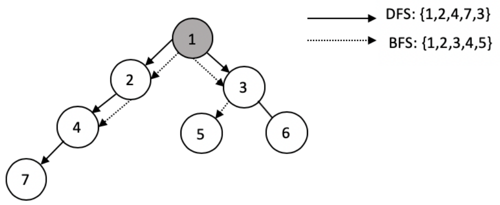

Figure 3: DFS and BFS for a *node* where number of walks = 4.

- Edge weights

P and Q are the probability that a node will retrace its path to the previous node or will go further to other undiscovered nodes, respectively. This probability depends on the *edge weight* (α), the normalized factor depends on the hyperparameters. Just like a word2vec skip-gram model where, for example, for the sentence "*I like horses*", the probability of the word "*horses*" depends on the occurrence of the words "*I*", "*like*" i.e. its surrounding; a node2vec graph also generates these directed subgraphs for the nodes in a particular walk[**?**]. As per the bipartite graph, the sample of nodes A is a union of patient nodes V and topic nodes T. Hence from equations 1 and 2:

$$A = V \cup T \tag{1}$$

5.4 Community Detection

The node embedding display some level of organization at an intermediate scale [40]. At this mesoscopic level, it is possible to identify groups of nodes that are heavily connected among themselves, but sparsely connected to the rest of the network. These interconnected groups are called communities, or in other contexts modules, and occur in a wide variety of networked systems [40].

Our ultimate goal is to identify unknown connections between patients and topics using the latent feature extracted. We use community detection method to find clusters in the high dimension feature space. The purpose of these *communities* is to incorporate similar patients and topics into regions defined by a boundary. These bounded regions are used to discover the unknown relationships between two patients or two topics and even connecting patients with topics which were not directly implied by the data.

In our paper, we aim to create communities which include both topics and patients(users) and observe the proximity amongst topics and patients. Apart from observing users in the community who have no relation with other users or topics, it will be interesting to observe the diversity of topics in one community. If two topics do not fall under the same 'Forum' (Figure 2) but are part of the same community, can help us determine the pattern and relationship between topics which you would not have been easily detected.

For detecting the communities we use k-means clustering and learn the clusters using Expectation Maximization (EM) algorithm. This method is divided into two steps - E and M steps. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster [6]. The objective of this

Table 3: This table contains the total number of users and topics before and after filtering in order to create a bipartite graph.

	Users	Topics
Original	94,393	140,000
New	43,680	100,018

method is to minimize cluster performance index, square-error and error criterion algorithm[24]. The algorithm tries to find K divisions to satisfy the optimal criterion. We manually went through a sample of clusters to verify whether our perceived outputs and the obtained outputs are similar in nature and modified the model hyper-parameters accordingly.

6 EVALUATION

In this paper, we present a novel approach to extract latent features from online breast cancer forums. Due to the lack of standard datasets or related work with the exact same objectives, evaluating our method can be challenging. We evaluate our method by building alternative models and demonstrate the improved performance of our method. We show that our design principles of building heterogeneous feature space (containing both patients and topics) has greater value compared to a purely text-based method and a third alternative where only patient interaction is modeled. We first describe the dataset and the network used for the experiments and results from a qualitative analysis.

We conducted several experiments to analyse the best configuration for the node2vec model, keeping in mind the large number of nodes as well as computation time. To obtain the optimal results without any bias we selected user-topic pairs from the original data at random and created the embeddings. We conducted our evaluation in two phases: 1. Qualitative Analysis - using a clustering algorithm on the embeddings and manually analyzing the obtained clusters 2. Quantitative Analysis - use a manually labeled reference dataset to evaluate our method using topic coherence and compare it with two other baseline models.

6.1 Experiment Setup

The original dataset from OBCF had 94,000 users with 140,000 topics. We eliminated topics with posts having posts made by greater 20 but less than 2000 users. This resulted in approximately 43,000 individual users and 100,018 topics for our model. We only used the User(patient) ID and no identity markers were used during this process. Detailed description of the dataset used in our experiments is summarized in Tables 3 and 4.

In accordance with our assumption, we draw a connection between a user and a topic if the user has posted in the said topic. We finally randomly sample 50,000 data rows to obtain *15,000 nodes* and *50k edges* (Table 4).

As our aim is to find out whether there is a relationship between the users posting across multiple topics. For example, if User v_1 and User v_2 have posted in a common topic T^* and User v_1 and User v_3 have posted in another common topic T^{**}, is it possible for User v_3 to be included in the community? Or due to the random nature of the walks is it possible for two or more topics who differ in their context to be included in the same community? Do they have a hidden connection and may have some commonalities that may help the users in their *road to feeling better?*

Table 4: Unique values of users and topics after randomly sampling 50000 data-rows from 'New' in Table 3 to avoid over-fitting of data.

Users	Topics
16431	9242

Table 5: Hyperparameters for node embedding model.

Parameter	Value
dimension	64
walk_length(l)	30
num_walks(r)	200
p	1.0
q	1.0

Table 6: Topic contents categorization for data points in the same community (Cluster- 01).

Topic	Context
F6 T779992	Managing Side Effects Breast Cancer & treatment
F83 T773037	Not Diagnosed but Worried
F44 T758994	Breast Reconstruction
F69 T784857	Chemotherapy Before, During and After
F78 T775441	Hormonal Therapy Before, During and After

The performance of the node2vec model depends on the values of the hyperparameters. We chose the default values of p and q i.e. *unity*. While a low q encourages outward exploration, it is balanced by a low p which ensures that the walk does not go too far from the start node[12]. Although the default value of dimensions d is 128, Grover et al. observed that the performance tends to saturate once the dimensions of representations reach around 100. After conducting multiple simulation experiments and keeping in mind the the large number of nodes in this model as well as the overall computation time, we decided empirically the value of number of random walks to be generated from each node (r), the number of nodes in each random walk (l) and the dimension d of each node is as given in (Table 5).

6.2 Qualitative Analysis

6.2.1 Objective. Our aim was to observe whether we obtain communities from the "Patient-Topic Network" which consist of topics as well as users and also display diversity in the topics in community. Our obtained communities contain users who have not posted in the topics included in that community but are still a part of it. This result is asserts our perceived output expectation as it depicts the probability of these users of facing similar issues that are discussed in the topics in that community.

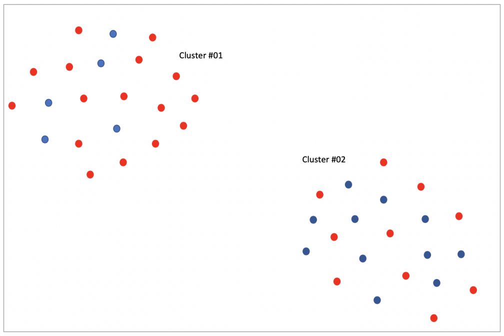

Figure 4: Representation of 2 communities out of 750 created from the node embedding model. Blue: Topics and Red: Users

6.2.2 Findings. We used k-means clustering algorithm for community detection. We used the elbow method of K-means to find the optimum number of clusters with a sliding range of 500-1500 clusters. Based on the outcomes of the elbow method we chose the number of clusters to be 750. Figure 4 is a representation of 2 out of 750 clusters obtained from the model and each cluster has about 20 data points. We can observed that "Cluster 01" consists of 5 topics and 15 users. The 5 topics in this cluster fall under 5 different forums (Figure 2). They have no common user between them but through this community we can get an idea about the relation between the users(patients) and topics which otherwise would not have been highlighted. From this community we can observe the diversity in the topics right from " Chemotherapy.." and "Breast Reconstruction" to "Managing Side Effects.." and "Hormonal Therapy- Before and After.." (Figure 2). We can observe that there is some relation between these topics as the topic names suggests, they range from discussing about chemotherapy and surgeries to the problems they faced due to the side-effects caused by them as well additional therapies required as a part of the treatment. The users(patients) in this clusters have some relation with either of the topics and hence with each other, can be advised on future difficulties well in-advance and the healthcare provider can create the treatment plans accordingly. In "Cluster 02' (refer Table 7) that the 11 topics in this community they are from 6 different forums. Topic F91 T792393, F91 T859005, F67 T796919, F93 T8605125 and F67 T27838556 has a user (patient) say v^* common between them. User (v^{**}) has posted in topics F5 T788278 and F5 T793169. Both v^* and v^{**} are also a part of this community. Treatment plans for users(patients) in this community can be created with respect to for example, the topics discussed by user([patient) v^*, as that patient has a diverse topic interaction and tracing their journey can help the other patients in this community.

6.3 Quantitative Analysis

In the previous section we presented a quantitative analysis of our model and its outcomes. In this section, we evaluate our approach quantitatively and by comparing the results against two baseline models.

Table 7: Topic contents categorization for data points in the same community (Cluster- 02). 2^* **refers to two other topics falling under the same category.**

Topic	Context
F91 T792393 + 2^*	Surgery - Before and After
F96 T835504	IDC Invasive Ductal Carcinoma)
F67 T796919 + 2^*	Stage III Breast Cancer
F93 T784857	General Comments and Suggestions
F16 T776398	For Caregivers, Family, Friends & Supporters
F5 T793169	Just Diagnosed

We constructed two baseline models – (1) a text based model and (2) user (patient)-based model. In the purely text-based model, we used Word2Vec embeddings [26] trained on the forum text. For the third model, we converted the data into a homogeneous network where there is only one type of node representing the patients. We use the Node2Vec [12] model to compute an embedding of the patients. Then we map this patient feature space into the topic space by replacing the patients with the topics they had directly interacted with. For example, if a patient v_i has interacted with the topics t_a, t_b, and t_c, in the embedded space we replace v_i with the three topics, each having the same embedding. If more than one patient has participated in the topic, the final embedding for that topic will be the centroid of all the patients' vectors who have participated in that topic. For all the three models, we have a mapping for forum topics and we evaluate our model by comparing the coherence of topics in each of these models. We also use a manually labeled dataset to represent the optimum topics and compare how well these models can replicate the reference set.

6.3.1 Results. Embedding methods provide a basis to obtain valuable insights on relationships between entities. Despite the huge popularity of these models, there are no well-defined metrics or methods that can be used to directly evaluate these models. Different methods have been used to measure the quality of embedding methods. Usually, these models are evaluated indirectly by measuring the performance of a downstream task (e.g. classification using the embeddings as features) or use human evaluators to judge the quality [35]. In our case, we do not have a well-defined downstream task but as the final step, we perform community detection. We use the quality of the communities to evaluate our overall methodology. Although our embedding framework is based on two variables – patients and topics, to be able to compare against the other baseline models, we only use the topic part of our model. We use two metrics for this evaluation, coherence of topics in the clusters and compare against a golden set of topic groups [17].

Coherence: Topic Coherence is a measure that looks into the degree of similarity between items in the topic and it is often used to measure the quality of the vectors in embedding methods [19]. These measurements help to understand how semantically interpretable the topics are. In our case, we extend this notion to measure the coherence of the topics (as defined in OBCF) to evaluate our method. There are numerous ways of measuring coherence statistically. Lau et al. [22] showed that normalized pointwise mutual information (NPMI) showed the most consistent correlation with

Table 8: Comparing the performance of the patient-topic network using Normalized Pointwise Mutual Information (NPMI) to measure coherence

	Coherence(NPMI)
Patient-topic network	0.481
User (patient) network	0.237
Word2vec	0.294

Table 9: The performance of the different methods in identifying communities with respect to the reference dataset. Here higher precision-recall values represent better identification of the communities as defined in the reference dataset.

	Precision	Recall
Patient-topic network	**0.643**	**0.588**
User (patient) network	0.428	0.314
Word2vec	0.507	0.422

a manually annotated test set, compared to other metrics, such as other variations of pointwise mutual information, Log Conditional Probability (LCP) [27] and pairwise distributional similarity [1]. We followed a similar approach and used NPMI to measure coherence in our case. Table 8 summarizes the results.

Comparison with Reference Dataset: Jones et al [17] manually categorized posts to identify actionable topic clusters from various online forums focusing on breast cancer, including OBCF. We used that manual set as a reference to evaluate our method using information retrieval evaluation metrics. Information retrieval systems are usually evaluated using two broad set of metrics – online and offline [25]. Online metrics measure the quality of the retrieved information using user engagement. In our case, we are evaluating our finding without the participation of any users, hence we need to use offline metrics. Among the numerous offline metrics, we selected precision-recall because we are not producing a ranked list of topics in this cluster. Thus, other metrics, such as, Discounted Cumulative Gain (DCG) or normalized DCG (NDCG) are not applicable here. Thus, we used precision-recall to measure the performance. The results are summarized in Table 9.

Both of these experiments show that the patient-topic network has performed better than the other variants. This demonstrates the strength of the heterogeneous network we used. The two variables in the network were able to setup a channel for better interaction and be able to extract richer features from the underlying data. As part of future work, we aim to design more experiments and further demonstrate the strength of this approach.

7 DISCUSSION AND FUTURE WORK

This paper presents a novel approach to identify features representing patients and the problems they face while undergoing treatment for breast cancer. We use an online forum to learn those features, where patients share their problems and difficulties brought upon by the diagnosis and the subsequent treatments. As patients discuss a variety of issues on these forums, we assume that these features will not only describe the medical issues a patient may face but also the personal, social, and professional problems that can affect a breast cancer patient's daily life. In this paper, we presented a model that mines this data and identifies those issues and extract features to represent the patients' state and the topics. This feature representation framework allows us to connect patients and topics and get a holistic view. That is, provide a base to know more about what a patient is facing currently or might face in the future. This is achieved by performing community detection on the high-dimensional feature space and identifying *similar* topics and patients. For example, if two patients v_x and v_y have shared their experience on a large number of common topics T_a and v_x has also talked about other topics T_b (i.e., $T_b \cap T_a = \emptyset$), our model is able to connect v_y with T_b, as there is a connection between v_y and T_b via v_x.

This paper presents the first step of the long-term goal of this work, which is to create a new knowledge base and integrate that with a physicians' existing workflow. This will help physicians and other healthcare providers to use this *additional* information and incorporate that to a patient's treatment plan. For example, before a physician prescribes a new treatment, they can estimate the effects of this new treatment on the patient's daily life. This will help the them to provide personalized/targeted treatments or be better prepared. However, this may introduce new challenges. For example, this new approach may raise privacy concerns. We can minimize the privacy risks by representing a patient and their state as a vector that represents their aggregated information. This vector, instead of representing the patient as an individual, will represent them with respect to other similar patients and the topics that are close to that patient. On the other hand, being able to place the patient within a larger perspective, the physicians can create a more holistic as well as personalized treatment plans. Our hypothesis is that without this new knowledge, physicians can only treat diseases and prescribe the treatment plan but fail to acknowledge a patients social and personal factors which may eventually become an obstruction in the patients road to *full* recovery.

We wish to further extend our study by involving clinicians and have them evaluate the perceived communities of topics and users. We also aim to integrate the final framework with EHR systems which will help the medical practitioners create a treatment plan for the individual patient based on their current diagnosis and what are the likely non-medical obstacles that may disrupt their recovery. We want to deploy this framework/tool as a web-based platform at various medical facilities across the state of Indiana for analyzing the clinical benefits of this study. We also want to extend this research hypothesis for other chronic diseases as well.

8 CONCLUSION

This paper presents a novel way to represent breast cancer patients using features extracted from an online health forum. We extract these features by converting the forum topics and the participation of patients into a bipartite network and used this network nodes for creating high-dimensional vectors. We observe that the newly constructed vectors can preserve the structure of the network as well as identify new relationships connecting similar patients or patients with similar topics, even when these relationships are not explicitly depicted in the data. We evaluate our model by showing the coherence of the new relationships and better performance compared to other similar methods.

REFERENCES

[1] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. 13–22.

[2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1568–1576.

[3] Danielle H Bodicoat, Minouk J Schoemaker, Michael E Jones, Emily McFadden, James Griffin, Alan Ashworth, and Anthony J Swerdlow. 2020. Correction to: Timing of pubertal stages and breast cancer risk: the Breakthrough Generations Study. *Breast Cancer Research* 22, 1 (2020), 1–2.

[4] Stacy M Carter, L Claire Hooker, and Heather M Davey. 2009. Writing social determinants into and out of cancer control: an assessment of policy practice. *Social science & medicine* 68, 8 (2009), 1448–1455.

[5] Nitesh V Chawla and Darcy A Davis. 2013. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine* 28, 3 (2013), 660–665.

[6] Immad Dabbura. 2018. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. https://towardsdatascience.com/k-means-clustering-algorithm-applications\protect\discretionary{\char\hyphenchar\font}{}{}evaluation-methods-and-drawbacks-aa03e644b48a.

[7] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi. 2019. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering* 2019 (2019).

[8] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. 2018. Bine: Bipartite network embedding. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 715–724.

[9] Lori J Goldstein, Raymond P Perez, Denise Yardley, Linda K Han, James M Reuben, Hui Gao, Susan McCanna, Beth Butler, Pier Adelchi Ruffini, Yi Liu, et al. 2020. A window-of-opportunity trial of the CXCR1/2 inhibitor reparixin in operable HER-2-negative breast cancer. *Breast Cancer Research* 22, 1 (2020), 1–9.

[10] William B Grant. 2020. Lower vitamin D status may help explain why black women have a higher risk of invasive breast cancer than white women. *Breast Cancer Research* 1 (2020), 1–2.

[11] Jeremy A Greene, Niteesh K Choudhry, Elaine Kilabuk, and William H Shrank. 2011. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of general internal medicine* 26, 3 (2011), 287–292.

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

[13] Robert A Hiatt and Nancy Breen. 2008. The social determinants of cancer: a challenge for transdisciplinary science. *American journal of preventive medicine* 35, 2 (2008), S141–S150.

[14] Fang Hu, Jia Liu, Liuhuan Li, and Jun Liang. 2019. Community detection in complex networks using Node2vec with spectral clustering. *Physica A: Statistical Mechanics and its Applications* (2019), 123633.

[15] Keyuan Jiang and Yujing Zheng. 2013. Mining twitter data for potential drug effects. In *International conference on advanced data mining and applications*. Springer, 434–443.

[16] Z. Jin, R. Liu, Q. Li, D. D. Zeng, Y. Zhan, and L. Wang. 2016. Predicting user's multi-interests with network embedding in health-related topics. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 2568–2575. https://doi.org/10.1109/IJCNN.2016.7727520

[17] Josette Jones, Meeta Pradhan, Masoud Hosseini, Anand Kulanthaivel, and Mahmood Hosseini. 2018. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. *JMIR medical informatics* 6, 4, e45.

[18] Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*. 43–47.

[19] Mohamad Abdolahi Kharazmi and Morteza Zahedi Kharazmi. 2017. Text coherence new method using word2vec sentence vectors and most likely n-grams. In *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*. IEEE, 105–109.

[20] Munui Kim, Seung Han Baek, and Min Song. 2018. Relation extraction for biological pathway construction using node2vec. *BMC bioinformatics* 19, 8 (2018), 206.

[21] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.

[22] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.

[23] Lu Li, Wei Wang, Shuo Yu, Liangtian Wan, Zhenzhen Xu, and Xiangjie Kong. 2017. A Modified Node2vec Method for Disappearing Link Prediction. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. IEEE, 1232–1235.

[24] Youguo Li and Haiyan Wu. 2012. A clustering method based on K-means algorithm. *Physics Procedia* 25 (2012), 1104–1109.

[25] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[27] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.

[28] Francois Modave, Yunpeng Zhao, Janice Krieger, Zhe He, Yi Guo, Jinhai Huo, Mattia Prosperi, and Jiang Bian. 2019. Understanding Perceptions and Attitudes in Breast Cancer Discussions on Twitter. *Studies in health technology and informatics* 2019 (08 2019). https://doi.org/10.3233/SHTI190435

[29] Laura Nyblade, Melissa A Stockton, Kayla Giger, Virginia Bond, Maria L Ekstrand, Roger Mc Lean, Ellen MH Mitchell, E Nelson La Ron, Jaime C Sapag, Taweesap Siraprapasiri, et al. 2019. Stigma in health facilities: why it matters and how we can change it. *BMC medicine* 17, 1, 25.

[30] Jungsik Park and Young Uk Ryu. 2014. Online discourse on fibromyalgia: text-mining to identify clinical distinction and patient concerns. *Medical science monitor: international medical journal of experimental and clinical research* 20 (2014), 1858.

[31] Jiajie Peng, Jiaojiao Guan, and Xuequn Shang. 2019. Predicting Parkinson's disease genes based on node2vec and autoencoder. *Frontiers in genetics* 10 (2019).

[32] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[33] Dina A Ragab, Maha Sharkas, Stephen Marshall, and Jinchang Ren. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* 7 (2019), e6201.

[34] Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. 2011. Scalable link prediction on multidimensional networks. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 979–986.

[35] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 298–307.

[36] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports* 9, 1 (2019), 1–12.

[37] Jennifer Y Sheng, Kala Visvanathan, Elissa Thorner, and Antonio C Wolff. 2019. Breast cancer survivorship care beyond local and systemic therapy. *The Breast* 48 (2019), S103–S109.

[38] Dongdong Wang, Nayden G Naydenov, Mikhail G Dozmorov, Jennifer E Koblinski, and Andrei I Ivanov. 2020. Anillin regulates breast cancer cell migration, growth, and metastasis by non-canonical mechanisms involving control of cell stemness and differentiation. *Breast Cancer Research* 22, 1 (2020), 1–19.

[39] Yang Yang, Nitesh Chawla, Yizhou Sun, and Jiawei Hani. 2012. Predicting links in multi-relational and heterogeneous networks. In *2012 IEEE 12th international conference on data mining*. IEEE, 755–764.

[40] Zhao Yang, René Algesheimer, and Claudio J Tessone. 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports* 6 (2016), 30750.

[41] Yongcheng Zhan, Ruoran Liu, Qiudan Li, Scott James Leischow, and Daniel Dajun Zeng. 2017. Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. *Journal of medical Internet research* 19, 1 (2017), e24.

[42] Enming Zhang. 2020. A Mixed-method Approach Towards the Understanding of Patient-generated Content on Social Media: A Case Study on Breast Cancer. Manuscript under review.

[43] Shaodian Zhang, Edouard Grave, Elizabeth Sklar, and Noémie Elhadad. 2017. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *Journal of biomedical informatics* 69 (2017), 1–9.