Annotating and Detecting Topics in Social Media Forum and Modelling the Annotation to Derive Directions-A Case Study

Athira B (athiramalu@gmail.com)
Cochin University of Science and Technology https://orcid.org/0000-0001-7830-2461
Josette Jones
BioHealth Informatics Deaprtment, IUPUI, Indianapolis
Sumam Mary Idicula
Cochin University of Science and Technology
Anand Kulanthaivel
BioHealth Informatics Department, IUPUI,Indianapolis
Enming Zhang
BioHealth Informatics Department, IUPUI, Indianapolis

Research

Keywords: Social media, Machine learning, Deep learning, Natural language processing, Breast cancer

Posted Date: February 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-132773/v2

Version of Record: A version of this preprint was published on February 27th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00429-7.

Annotating and Detecting Topics in Social Media Forum and Modelling the Annotation to Derive Directions-A Case Study

Athira B
Dept. of Computer Science
Cochin University of Science and Technology
Cochin-22, India

Josette Jones
BioHealth Informatics Department
IUPUI, Indianapolis
IN 46202, United States

Sumam Mary Idicula
Dept. of Computer Science
Cochin University of Science and Technology
Cochin-22, India

Anand Kulanthaivel
BioHealth Informatics Department
IUPUI, Indianapolis
IN 46202, United States

Enming Zhang
BioHealth Informatics Department
IUPUI, Indianapolis
IN 46202, United States

Corresponding author:
Athira B
Department of Computer Science
Cochin University of Science and Technology
Cochin-22
India
Email: athiramalu@gmail.com

Abstract

The widespread influence of social media impacts every aspect of life, including the healthcare sector. Although medics and health professionals are the final decision makers, the advice and recommendations obtained from fellow patients are significant. In this context, the present paper explores the topics of discussion posted by breast cancer patients and survivors on online forums. The study examines an online forum, Breastcancer.org, maps the discussion entries to several topics, and proposes a machine learning model based on a classification algorithm to characterize the topics. To explore the topics of breast cancer patients and survivors, approximately 1000 posts are selected and manually labeled with annotations. In contrast, millions of posts are available to build the labels. A semi-supervised learning technique is used to build the labels for the unlabeled data; hence, the large data are classified using a deep learning algorithm. The deep learning algorithm BiLSTM with BERT word embedding technique provided a better f1-score of 79.5%. This method is able to classify the following topics: medication reviews, clinician knowledge, various treatment options, seeking and providing support, diagnostic procedures, financial issues and implications for everyday life. What matters the most for the patients is coping with everyday living as well as seeking and providing emotional and informational support. The approach and findings show the

potential of studying social media to provide insight into patients' experiences with cancer like critical health problems.

Keywords: Social media; Machine learning; Deep learning; Natural language processing; Breast cancer

1. Introduction

Social media is a rich resource for gathering information, even on personalized topics such as health, wellness, and decision-making. People gather in discussion groups to verbalize concrete health challenges and obtain input from peers. Consequently, the role of the patient has transformed from that of a passive receiver of health information to a provider of knowledge and information. This transformation has spawned an emerging field of social network explorations, the patient-centric social network [1]. The users of such a social network are patients (or caregivers) with certain health conditions or symptoms seeking possible solutions. Examples of these online health communities (OHCs) are Patientslikeme, Dailystrength, and Inspire [2]. They center on medication experiences, treatments, symptoms, diagnoses, and nonmedical information, such as family and personal dilemmas; people seek and share stories of various challenges and even look forward to receiving emotional support. Hence, these groups provide great solace to these patients in terms of informational as well as emotional support.

Although breast cancer is the most common cancer among women globally, its prognosis is steadily improving owing to early diagnosis and timely and effective treatment. Nevertheless, managing the disease and maintaining the quality of the daily life of the patient is worth exploring. Several studies have been performed on cancer survival; however, there is scant knowledge of the “experience” of living with cancer. The factors associated with the unmet post treatment needs of cancer survivors are fear of recurrence, lack of up-to-date information, and diminished quality of life [3]. Other support needs of women with breast cancer were discussed in [4]. The prevalence of depression, anxiety, and distress among long-term cancer survivors and the psychosocial factors that follow diagnosis affect long-term survival [5–6]. These factors show the potential of studying the experience of patients and thereby providing insight into the necessity of primary care.

Identifying topics of discussion in online health communities (OHC) is critical to study patients' experience. Nevertheless, it is challenging to read and understand the extensive material published on OHCs because they are

typically diverse and domain-dependent. In one of the previous studies [7], topic modeling on a breast cancer community forum, Breastcancer.org, was performed in an unsupervised manner to obtain a general idea of the topics of discussions. They have identified 4 clusters of topics as 'symptoms &diagnosis', 'treatment', 'financial' and 'Friends & Family'. To obtain deeper insight into the topics and what matters most about them, the current study used 3 million posts from the same OHC, Breastcancer.org. and manually annotated 1000 posts with different topics of discussion. Subsequently, a machine learning model for the base classifier was created for topic classification. The greatest challenge of the current study was the limited size of labeled posts. It was not adequate to cover all the features of our large 3 million datasets of unlabeled posts that appeared during the prediction task. The accuracy and performance of machine learning models, especially deep learning models, depend on a large amount of training data. Hence, a semi-supervised learning technique was used to increase the volume of the training data. Finally, a deep learning model was used to classify the topics of discussion. The topics in the current study provide an excellent pathway to more personalized healthcare that incorporates clinical and nonclinical observations with patient-generated data.

In creating the model, colloquial and everyday language, shorthand terms with nonstandard grammar, and typographical errors posed several challenges. Furthermore, the necessity of including and interpreting specific informal terms precluded the possibility of utilizing any standard natural language processing (NLP) platform. The following is a typical post: "VR - agreed! I got it based upon your recommendation. lol!!!!". Hence, it was essential to translate this noisy text into a regular and organized document. Another major challenge was generating labels from the vocabulary of these typical posts. For instance, the post "Hi I have some BC stuff I don't need any more and would like to donate to someone in need" is annotated as "support for offering material as donation" (SUP).

The main objective of the current research is to obtain better insight into the topics of discussions among patients by studying the communication messages and thus draw conclusions about what matters to patients. To accomplish the main goal following distinct approaches were proposed:

(i) Develop an effective method for standardizing the text so that it is compatible with natural language tools.

(ii) Build an effective method to increase the volume of the training data for the benefit of deep learning models.

(iii) Build an efficient deep learning model to represent the 3 million posts and learn the semantic characterization of the correct annotations.

iv) Finally obtain the knowledge of overall prevalence of topics on the breast cancer community

First, normalizing and cleaning text to enable classification models to minimize information loss was developed. A base classifier was then built with 1000 labeled posts, followed by a semi-supervised technique to increase the size of the training data. With the increased volume of training data, a few deep learning models CNN, LSTM, and BiLSTM, were implemented, and the efficacy of these algorithms in detecting meaningful labels was compared. Ultimately, the general prevalence of topics in the entire community forum was analyzed.

2. Related Work

Most of the related research on these kinds of online forums shows that the primary intention is to provide support in terms of information, emotional support, recommendations from the expertise of other patients, and, above all, feeling that "I am part of a group or community" [8]. The research group in [9] compared the adverse drug reactions (ADRs) mentioned in social media with a drug information database and pointed out that social media is a critical source of information from a patient perspective. The studies in [10] and [11] clarified the significance of identifying a group structure in the same way as likely consumers are identified. The researchers in [12]-[15] focused on the smoking cessation community, the temporal analysis of online cancer forums, the social support effectiveness of an online breast cancer support community, and understanding depressive symptoms and psychosocial stressors from Twitter data, respectively. In online cancer groups, relatively little research has been reported based on posts from the user community. Positive and negative opinions of patients on using the drug 'Tamoxifen' were explored in [16]. Elhadad et al. [17] identified three words for the semantic forms describing situations arising during treatments: drugs, symptoms, and side effects. Yang et al. [18] performed another significant research area in OHC, where an effective thread recommendation is proposed to locate the information of interest.

Several published studies have concentrated on processing posts to elicit meaningful information. Liu et al. [19] stressed the need for an efficient NLP tool to improve social media text quality. A novel algorithm for

enhancing the preprocessing phase was proposed in [20]. Lee et al. [21] dealt with the problem of colloquial language in social media posts and described a novel approach for mapping medical concepts to layman texts. A rule-based approach to refining the text in posts was explained, and the rules of the database and the classification system were presented by Clark et al. in [22]. A recent work by Conway et al. [23] specifically conducted a literature review of NLP tasks on the consumer-generated text in social media for public health. They identified various research directions for social media and stressed the lower usage of modern machine learning approaches in this area. Therefore, the work reported here utilizes vectorization methods such as TF-IDF, word2vec, and doc2vec [24] to process the text and uses several classifiers to create machine learning models. Bidirectional Encoder Representations from Transformers (BERT) and Embedding from Language Models (ELMO) are recent developments [25][26] in vectorization methods for generating context-dependent word embedding vectors to perform composite NLP tasks. In a very recent study [27], doc2vec models and bag-of-words representation were used by authors to extract the features and classified the support given in the posts in an OHC. Yet, in another study [28], word2vec and NER models were used to evaluate the quality of answers in OHC.

A recent study by [29] classified the topics from a cancer community forum by observing the sentences in individual posts. A deep learning model, CNN, was used in the study to classify posts using word vectors. The labeled data samples for each class were the sentences rather than the posts themselves. Therefore, the training data for the CNN was large, which is essential for the performance of the deep learning model. However, in the current study, labeled data samples were insufficient to build the classification model for the big dataset size used. The literature has shown that a lack of labeled training data can be addressed by semi-supervised learning techniques [30]. This learning technique's primary focus is to merge labeled data and unlabeled data to create better learners. The semi-supervised approach is usually formulated on top of the supervised method, i.e., a supervised classifier is used as a base classifier [31] to predict the labels for unlabeled data. Self-training and co-training are the two types of approaches in semi-supervised learning techniques. In self-training, a base classifier built on a small set of labeled data is used to predict the labels of unlabeled data, and in co-training, two classifiers built on two disjoint sets of labeled data are used to predict the labels of unlabeled data [32]. Self-training is one of the simplest techniques in

which the learner continues to label unlabeled data and retrains itself on enhanced labeled training data [33]. However, one disadvantage of self-training is that if a particular data instance receives an incorrect label at any iteration, this error will propagate to the subsequent training phases. There are some variants of self-training that can reduce this error. One such approach is to use only high-confidence predictions from the initial prediction model. A base classifier is needed for this technique to predict the probability score for prediction, and this probability score is used to filter the labeled retraining data. A study in [34] achieved a further improvement in this area. More confident labels are predicted by self-training with a look-up table and thus significantly reduced the propagated error.

It was noted from the literature review that while there were many studies in various dimensions in OHCs, only a few studies were reported in cancer community forums to identify the topics of discussion. The size of the labeled data samples is a key challenge in the topic classification task. From the reported performance evaluation of the related study in [29], it was observed that the size of the training samples and the feature vector representation of words are essential for the better performance of a deep learning model to classify a large amount of data. Therefore, the key goal of the current study is to use a better self-training approach to increase the size of training data and use an efficient vector representation of words to reflect semantic characterization. Towards the end, the study assessed the efficacy of various deep learning algorithms to classify community posts.

3. Methodology

3.1 Data Source

The data set used in this study consisted of more than 3 million anonymous posts from the Breastcancer.org site. Permission to collect these data was granted by the site administrators. The site permits its users to publish medical details as part of their user signatures. The medical history of users is, by default, private according to the site's policy. However, users may choose to publish them in whole or in part by modifying their privacy settings. Once a user decides to publish a specific medical history, it becomes part of the signature to every post they make and will be visible to anyone reading the post. Medical histories from signatures were scraped and stored in a database. Python's beautiful soup package was used for the data collection. In total, the database has 3,312,674

discussion postings written by 78,553 users. A subset of 10 of the 80 forums from the site was used. Each forum focuses on different aspects of breast cancer management such as insurance, new patients, material support, invasive nature, etc. Thus, a total of 1000 posts were collected through Nth member sampling [27]. In the present work, N=5, i.e., every 5th post from each of the forum was taken, and a total of 1000 sample posts were collected. A combination of forum titles and random post-selection were utilized to determine the ten forums. The forums chosen for this study, the sample posts, and the manually annotated labels are listed in Appendix 1, which also contains various sublabels

3.2 Manual Annotation

As in [35], a qualitative content analysis formed the basis of creating a bottom-up thesaurus of user communication topics. Furthermore, six annotators, graduate students in health informatics with clinical backgrounds, were trained on the annotation strategies. They were divided into two groups of three members each. Two groups were asked to annotate the same samples of text in parallel. First, the two groups read each post and understood its meaning, and then they created a code (label) to capture the patient's perspective and priority. Then, their annotations were compared, and a kappa score (agreement score) was computed. A kappa score of 0.63 was achieved. The labels were later reviewed by the team and validated by a clinical expert on the team. The study in [36] used an analogous methodology to categorize the everyday problems faced by patients. The labels given to the topics, their frequencies, and their definitions are shown in Table 1. The given labels address many aspects of everyday health problems for patients with breast cancer.

Table 1: Frequency of topics

Topic	**No of posts**	**Definition**
SUP (Support)	211	Support given or received by patients in terms of information as well as emotional support
CTXF (Contextual Factors)	184	Entities encountered during everyday life, specifically by ill patients and survivors.
TXS (Treatment Stories)	154	Discussing treatments
ADE (Adverse Drug Effects)	109	A negative reaction a patient experienced and must cope with; attributed to a drug
DXS (Diagnosis Stories)	122	Discussing diagnostic stories and histories

FIN (Finance)	115	Financial matters, including insurance
CLIN (Clinic/Clinician Information)	105	Information about clinics or clinicians

Although these seven labels included various sublabels, as shown in Appendix 1, only the parent labels in the hierarchy were considered in this study for classification. Besides these 1000 posts, a few 20 posts containing emotional elements such as greetings, gratitude, and disappointments. However, owing to the small samples, these posts were not considered for classification. A sample post in each of the topics is shown below.

1. Class 1: ADE - “stopped because of allergic reactions - one of which being severe folliculitis that lasted three weeks”.
2. Class 2: CLIN - “dr. jamie escallon is a great breast surgeon out of womens college and princess Margaret”.
3. Class 3: SUP - “do you still need those headscarves? i have a couple of extra pretied bandanas I could send you that i got off of myheadcoverings.com. would those interest you?”
4. Class 4: CTXF - “i have an old drug co. freebie that i now hold together with rubberbands and it holds 12 days of am and pm meds in my carry on. i also pack my daily prescription meds in my luggage or in a little shopping bag if in the car. i too want to be prepared for an extended stay in an emergency”.
5. Class 5: DXS - “hi i'm at stage iiia dianosed june 1 2009. had lumpectomy in july. tumor was 2cm x 2cm x2.5cm. 7 out of 16 lymph node involvement. has small ilc tumor in other breast w/ no lymph node involvement. had scant trace of cancer cells left in my right chest wall (no clean margins)”.
6. Class 6: FIN - “itsjustme...i am still waiting for the actual bill to come. at this point i have no idea who sent bits of my body to this lab. when i get the bill i hope to see the information i need to resolve this. especially after reading your post it does not make sense that everything else is covered and then there would be this strange lab showing up with such a large fee”.
7. Class 7: TXS - “i just hit the 5 year anniversary of my diagnosis this month. i was stage 3a grade 3 with 6 of 9 positive nodes. i had tac chemo rads hysterectomyabilateral mastectomy diep reconstruction 1 year of tamoxifen (before the hysterectomy) and i'm in year 4 of arimidex. i'm also in the clinical trial for biphosphonates (i'm in the zometa arm of the trial)”.

3.3 Overall Approach

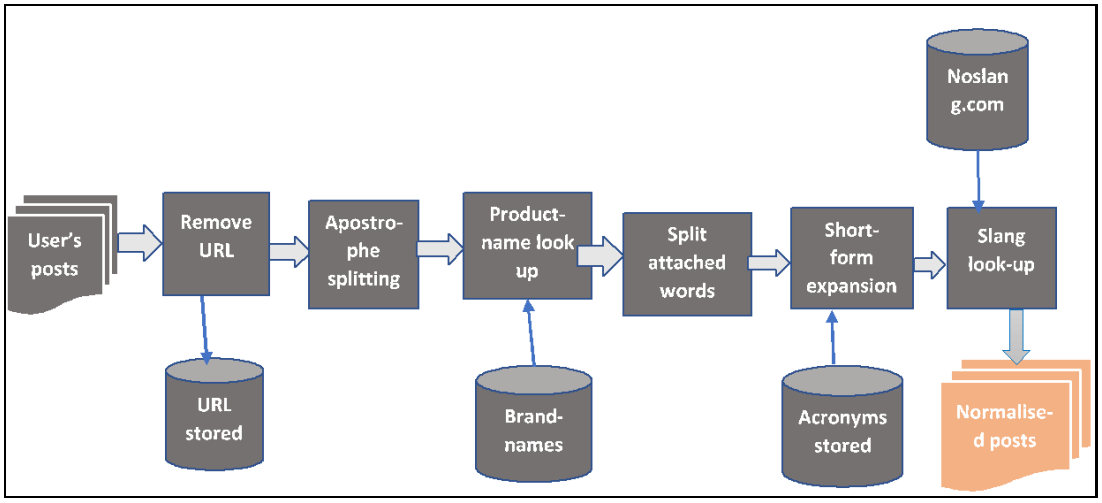
The manually labeled posts were divided into two categories: training and testing. The posts were normalized and cleaned, after which they were converted into vectors of real numbers using TF-IDF, word2vec, and doc2vec techniques. Next, a base classifier was built with 1000 labeled posts, and the results obtained were

compared with those obtained using the BERT feature vector. In the second stage, a self-training technique was implemented to increase the size of the training data. Self-training with a lookup list was used to optimize the training data. Finally, with enhanced training data, three deep learning algorithms, CNN, LSTM, and BiLSTM, were implemented, and the results were compared to classify all 3 million posts.

3.4 Normalizing and Cleaning the Text

To address the problems arising from unstructured text and misspellings, the sequence in Figure 1 was implemented.

Figure 1. Framework for Normalizing and Cleaning



1. Removal of URLs: The URLs and hyperlinks were eliminated based on regular expressions in the text data, including comments and feedback. All the URLs were deleted and replaced with the word "url." The removed url address was stored in a database where it could be referenced in cross-linking for future use.
2. Stripping of apostrophes*:* Apostrophes increase the chances of disambiguation. For example, "it's" is the contraction of "it is" or "it has". All clitics were converted into a standard form. A look-up table of all possible forms was used.
3. Product name lookup*:* Users used the names of many products to express their opinions, which were all domain related. However, the current study consciously adhered to a practice of not endorsing or disparaging products of any type. A dictionary of the names of all the domain-related products that were noted during the manual

annotation process was created in order to avoid the use of any product brand names. These terms were replaced by the expression “product_name.”

4. Split attached words*:* People tend to use very informal text on social media. Some posts were accompanied by several hashtags, such as RainyDay and PlayingInTheCold. Using regular expressions, these entities were split into basic words.
5. Standardizing words: The users used many short forms, such as chemo for chemotherapy, rads for radiation, and BC for breast cancer. Thus, a dictionary of all the domain-related acronyms and their expanded forms was created.
6. Slang lookup*:* As stated earlier, social media consists of many slang words. It is necessary to translate these words into standard words. For instance, “lol” is translated into “laughing out loud”, and “omg” is translated into “oh my god.” The internet slang was translated with reference to noslang.com.

Next, traditional preprocessing tasks, such as the removal of punctuation marks and deletion of stop words like “the,” “for,” “a,” and “an,” were conducted. Lemmatization, the reduction of the inflected form of a word to its root, was also performed. For instance, “waiting,” “waited,” and “waits” were reduced to the root form, “wait.” The final step was to represent the words as bigrams.

Next, the bigram terms were converted to a proper representation in terms of vectors, called feature vectors, to facilitate the creation of the classification model. To create the feature vectors, TF-IDF and word embedding techniques such as word2vec, doc2vec and BERT were used. Apart from the steps mentioned above of normalization, the other text inconsistencies such as 'typos' were well managed at word vector formation using these techniques. Word embeddings are essentially mathematical representations of meanings based on word distribution. Word embeddings are created from the principle of "You shall know a word by the company it keeps," i.e., words that are in similar contexts have similar vectors. In this sense, very regular misspellings are needed to get different vectors in the vector space approach. The details of these techniques are described in Appendix 2.

3.5 Building the Base Classifier

First, the labeled data were split into training and testing sets in 80% and 20% ratios, respectively. Of the 1000 posts available, 800 were used for training, and the remaining 200 were used for testing. The sequence of the training vectors Xn, where n is 800 (corresponding to 800 posts), produced from the above step, were fed to a KNN classifier. A KNN classifier has seven clustering centers, one corresponding to each label. The Euclidian distance measure was used to ascertain the membership of each class. Subsequently, a neural network with the following specifications was implemented: 100 input layers, one hidden layer with 50 neurons, and seven output layers; the logistic function used was Tanh. The network was trained using backpropagation. The classification was also attempted with an SVM with a radial basis function kernel and gamma=10^{-3}. All the above classifier functions were implemented using the Scikit library in Python 3.6, and 10-fold cross-validation was performed. For all three models, the resultant F1-score was less than 65%. The results are tabulated in Table 5 in the Results section.

An ensemble neural network (ENN), which is robust to small data sizes [40], was used to improve the classification performance. Compared with the classifiers mentioned above, the ENN yielded better results. The data imbalance problem was attempted to address by using oversampling and under-sampling techniques with the synthetic minority oversampling technique (SMOTE); however, this did not yield promising results. According to the literature, an ensemble of neural networks can solve imbalances in a data set and can select the appropriate representative data sample from each class effectively [41]. ENN was implemented with multiple hypotheses for different combinations of the training and testing data. A hypothesis was generated from input X to output Y, where X is feature vectors of posts and Y is the corresponding label. The training and testing data vectors were processed concurrently and were randomly chosen from the training data set of 800 elements. The basic idea is that the samples that were misclassified by the current hypothesis were used as the training sample to formulate the subsequent hypotheses. Likewise, the generation of hypotheses continued until the misclassification stabilized at 20% and resulted in a set of seven hypotheses. The remaining 200 test data vectors were fed to the seven hypotheses simultaneously. The best classification was obtained through weighted majority voting, and Figure 2 shows the algorithm for implementing the ENN.

Figure 2: Algorithm for the ENN

Input: 1) Data set D with n samples $((X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n))$ where $X_1, X_2 \ldots X_n$ are n data points in the data set and $Y_1, Y_2 \ldots Y_n$ are the n corresponding annotated labels.

2) $W_i = 1/n$ where the initial weight for each data points, $i=1$ *to* n

3) A classifier

4) 'm' number of iterations

Output: Combined hypothesis, $H_{final} = \sum_1^m h_j(Score_j)$

while $j=1, 2 \ldots m$ or classifier's error stabilize

1) Randomly choose TR_j and TE_j from data set D. Create hypothesis h_j: $X \rightarrow Y$ on $TR_j + TE_j$
2) Call weak classifier, provided with TR_j and Calculate the classification error on TE_j, $\varepsilon_j = \sum_{i:h_j(x_i \neq y_i)} W_i(j)$ on TE_j
3) If $\varepsilon_j > \frac{1}{2}$, discard h_j and go to step 1, otherwise, normalize error, $\beta_j = \varepsilon_j / (1 - \varepsilon_j)$
4) Calculate $score_j = \sum_{i:h_j(x_i=y_i)} \log\left(1/\beta_j\right)$, for $\forall$ Y
5) Update weight: $W_i(j+1) = W_i(j)$ x $\begin{cases} 1, \text{if } h_j(x_i \neq y_i) \\ \beta_j, \text{otherwise} \end{cases}$

$Combined\ Score_j = \underbrace{\text{argmax}}_{y \epsilon Y} score_j$

Inputs to the algorithm were feature vectors $X_1, X_2 \ldots X_n$ of 800 training posts, and corresponding labels $Y_1, Y_2 \ldots Y_n$ where n=800. Initially, all the data points were given equal weights, $W_i = 1/n$ where $i=1$ *to* n, so they had an equal initial probability of being chosen as samples for training and testing. The training (TR_j) and test sets (TE_j) were then randomly selected from the training data to generate hypothesis h_j. Build a weak classifier with TR_j and test with TE_j. The error, which was the sum of the weights of the misclassified instances, was measured using:

$$\varepsilon_j = \sum_{i:h_j(x_i \neq y_i)} W_i(j). \quad (1)$$

where ε_j is the error at j^{th} hypothesis, $W_i(j)$ is the weights of i misclassified instances $(h_j(x_i \neq y_i))$ at j^{th} hypothesis. The data vectors misclassified by h_j from TE_j were fed into the next hypothesis, h_{j+1}, as training data TR_{j+1}. If the error ε_j is more than 50%, then the corresponding hypothesis was discarded, and the new hypothesis was formulated; otherwise, the error was normalized. It was interesting to note that no more than three hypotheses were rejected for the data set used.

The error was normalized, denoted as β_j, and calculated as:

$$\beta_j = \varepsilon_j / (1 - \varepsilon_j). \quad (2)$$

A hypothesis score for the current hypothesis (h_j) was then calculated as follows [42]:

$$score_j = \sum_{i:h_j(x_i=y_i)} log\ (1/\beta_j). \quad (3)$$

$score_j$ was based on the correctly classified instances for each class. And the value was depending on the number of instances that were correctly classified ($h_j(x_i = y_i)$) according to the specific class.

Then, the weights of the instances were updated. The weight of each misclassified instance was unchanged, whereas the value of β_j reduced the weights of the correctly classified instances; hence, the probability of misclassified instances being included in the training sets increased.

$$W_i(j+1) = W_i(j)\ x \begin{cases} 1, if\ h_j(x_i \neq y_i) \\ \beta_j, otherwise \end{cases} \quad (4)$$

Where $W_i(j+1)$ is the weight for instance i in $j+1$ hypothesis.

Thus, the algorithm was iterated until the classification error stabilized. Finally, the combined score ($Score_j$) was determined for each class as the maximum voting score from all the scores ($score_j$) considered:

$$Score_j = \underbrace{argmax}_{y \epsilon Y}\ score_j \quad (5)$$

The final hypothesis H_{final} was developed as follows: the $Score_j$ for each class was mapped to the best hypothesis (h_j) for that class, and all the best hypotheses for each class were eventually combined to create the final hypotheses for the classification model.

$$Combined\ hypothesis,\ H_{final} = \sum_1^m h_j(Score_j) \quad (6)$$

3.6 Self-training to Increase the Training Data Size

The training data posts had to be enhanced to label all of our data posts from the breast cancer site. As stated in the Related work section, a semi-supervised learning technique, self-training with a look-up list, was used here. A base classifier is implemented and used for the unlabeled data instance's prediction. The predicted label of the unlabeled data instance is not added directly to the training set at the first time. Instead, it is added to a look-up list. If it receives the same label again in subsequent iterations, then the data instance is added to the training set with the received label.

First, a neural network classifier was built with the combined hypothesis H_{final} from the above algorithm. Then, it was trained with a self-training technique for predicting the rest of the unlabeled data posts. The rationale behind using the neural network was that compared to the KNN and SVM methods, the neural network yielded a better result in the first phase of building the base classifier. Approximately 100,000 posts, termed U, were collected from the unlabeled posts for the current process. After preprocessing and feature vector generation, a sample of 200 posts, termed u_i, was fed to the classifier for prediction at each iteration. Since the BERT vectorization technique yielded a slightly improved result, it was used here to create the vectors for each post. The probability score from the classifier was noted. The posts that received a label y_i with a probability score greater than a threshold value k was added to the look-up list. The process was iterated through the following steps:

1) If, for a label y_i of an unlabeled instance u_i, the probability score was $\geq k$, then u_i was added to a lookup list with label y_i.
2) In the subsequent iteration, if u_i obtained the same value y_i as the label according to the look-up list, then u_i was added to the training set.
3) To avoid data imbalances, an equal number of labeled instances from every class was added to the training instances for the next iteration.

Thus, with more training data posts in each iteration, the model was subsequently learned. Out of 100,000 unlabeled data posts, 74,324 posts were ultimately labeled and added to the training data set. The efficacy of this technique was evaluated with different training sizes of labeled posts of 40%, 50%, 60% and 70%. In each of these cases, 60%, 50%, 40% and 30% of the labeled posts were used for testing. Three different values of k, that is, 0.85, 0.90, and 0.95, were also used for experimental purposes. The results are shown in Table 6 in the performance comparison of the self-training classifiers in the Results section.

3.7 Deep Learning Classifiers

With enhanced labeled data posts, three deep learning-based classifiers were implemented as the final classification models so that all 3 million posts could be classified more efficiently. The three models, that is, CNN, LSTM, and BiLSTM, were implemented with the BERT embedding technique, and the efficacy of these algorithms

was compared. With the CNN, two convolution layers and three filters of stride three were used. After preprocessing, each word of the posts was converted to a BERT vector. The word vectors were fed to the convolution layer as input, and create feature vectors using filters. The filter is a matrix filled with weights that perform convolution operation on word vectors to generate feature vectors. Thus, by sliding over the sentences' word vectors, each filter performed a convolution operation to create a feature map. This feature map was fed to the next layer, the max-pooling layer, whose function is to reduce the feature map's size. In the max-pooling layer, the largest score of the feature map was taken. Finally, a fully connected layer with a softmax function classified each post. The max-pooling layer's output was fed to a fully connected layer, which is similar to a multi-layer perceptron neural network consisting of neurons connected between different layers. From this stage, the classification process begins. The fully connected layer's output was numerical values for each class, and the maximum of these values was calculated by the softmax function as the probability for each category. The class that achieved the highest predicted probability is the predicted class.

Another deep learning model used was LSTM, a variant of a recurrent neural network, which can classify sequences of data, such as text, more efficiently. LSTM consists of many cell units called memory units that can either remember or forget information. Each cell consists of an input gate, an output gate and a forget gate that control the information in the cell. These gates collectively update the current memory cell and current hidden state. The input gate, i_t determines how much data is stored in the current memory cell, output gate, o_t, determines how much data is to be output, and forget gate, f_t determines how much information is to be thrown out. After converting each word with the BERT vector, the sequence of word vectors was fed to each cell unit. At each time step t, the current memory cell c_t and current hidden state h_t are:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$h_t = o_t \odot tanh\,(c_t)$$

where $\odot$ is element wise multiplication, c_{t-1} is previous cell state, *tanh* is hyperbolic tangent function that has value between [-1,1] and allows the cell state to forget memory.

The input gate, i_t output gate, o_t and forget gate, f_t transition functions are as follows:

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

$f_t = \sigma\,(W_f \cdot [h_{t-1}, x_t] + b_f)$

$o_t = \sigma\,(W_o \cdot [h_{t-1}, x_t] + b_o)$

where σ is sigmoid activation function that has value between [0,1], x_t is current input word vector, h_{t-1} is previous hidden state, b_i, b_f, and b_o are biases and W_i, W_f and W_o are weights. The above equations were repeated for all the memory cells. Finally, the softmax function was applied to the last cell output to obtain the highest probability of the class.

The third deep learning model implemented was BiLSTM due to its bidirectional nature in capturing text sequence information. BiLSTM effectively obtains information from the future and past sequences of words by forming two LSTM networks layers. Past information can be extracted by providing the words in a sentence from left to right to the first LSTM network, and future information can be extracted by providing the words in a sentence from right to left to the second LSTM network. The output values from the left LSTM units and right LSTM units were computed as stated above in equations (6) and (7), and were then concatenated and provided to the softmax layer to classify the posts.

3.8 Evaluation Metrics

The performance of the various classification algorithms reported was evaluated using the metrics of precision, recall (sensitivity), and F1 score, which can be formally defined as follows:

1) Precision is the ratio of correctly predicted instances to the total predicted instances.

   Precision = TP/TP+FP, where TP = true positives (correctly predicted), FP = false positives (wrongly predicted), and TP+FP = total predicted.

2) Recall is the ratio of correctly predicted instances to the actual instances of that category.

   Recall= TP/ TP + FN, where FN= false negatives and TP+FN= actual instances

3) The F1-score represents the balance between precision and recall. The F1-score is the harmonic mean of precision and recall given by $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$

4) Accuracy is the ratio of the number of correct predictions to the total number of predictions made.

   Accuracy= (TP+TN)/(TP+TN+FP+FN)

4. Results

4.1 Normalizing and Cleaning Results

To evaluate the performance of the normalizing and cleaning steps, two performance measures, the word frequency and lexical richness were compared with and without the proposed normalizing and cleaning framework on manually labeled posts. Table 2 displays the lexical richness of the forum before and after normalizing and cleaning. Initially, the distinct tokens (words) composed only 21% of the forum; following normalization and cleaning, this increased to 44%. Thus, the new preprocessing algorithm doubled the lexical richness.

Table 2. Lexical richness of the forum before and after normalizing and cleaning

	Before normalizing & cleaning	**After normalizing & cleaning**
Total no. of tokens in the forum	114,345	41279
Total no. of unique tokens	24933	18184
Lexical richness	24933/114345=0.218	18184/41279=0.440

Figures 3 and 4 show the frequency distribution of the top 50 tokens on the posts before and after normalization and cleaning. Figure 3 indicates the noise as the tokens ":" and ".."; stop words were the most repeated tokens. According to Figure 4, the top tokens in the forum were "url," "free," "need," "product name" and "breast_cancer," because most of the active participants on the forum were breast cancer survivors who were willing to donate several treatment items, such as wigs, scarves, breast binders, and medicine. This result was further corroborated by the fact that "SUP" was a top-level topic. The full length of each URL was substituted with the word "url," and the names of the products were replaced by the term "product-name". The resulting bias of the graph towards more frequently occurring tokens demonstrated the effectiveness of removing the long URL addresses and product/brand names from the text. There was no obvious information loss in the entire process.

Table 3 shows the performance of the models before and after normalizing and cleaning. The ENN model improved significantly after normalizing and cleaning. This result underscores the importance of the proposed normalizing and cleaning steps used.

Table 3. F1-scores of models before and after normalizing and cleaning phase

Accuracy

	Before normalizing and cleaning	After normalizing and cleaning
ENN	0.664	0.721
Neural network	0.532	0.648
SVM	0.525	0.634
KNN	0.512	0.625

Figure 3. Top 50 tokens before normalizing and cleaning

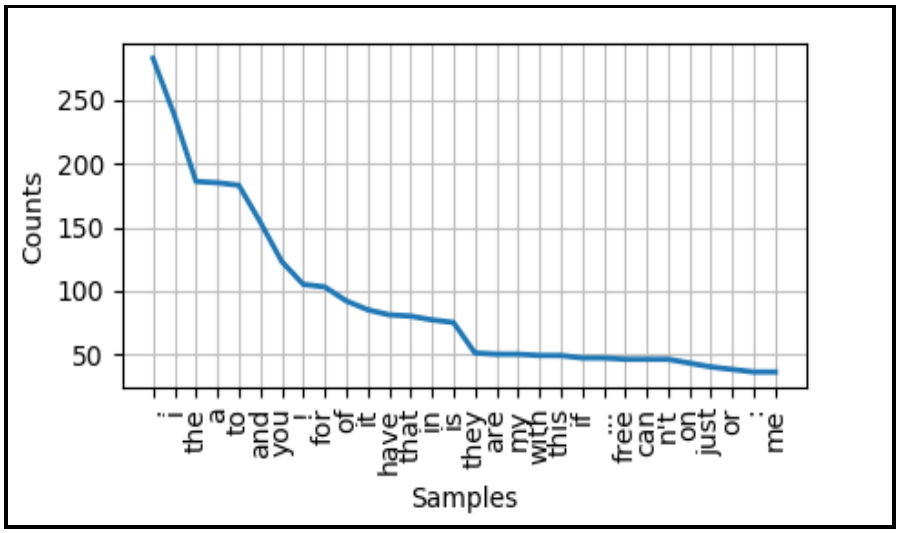
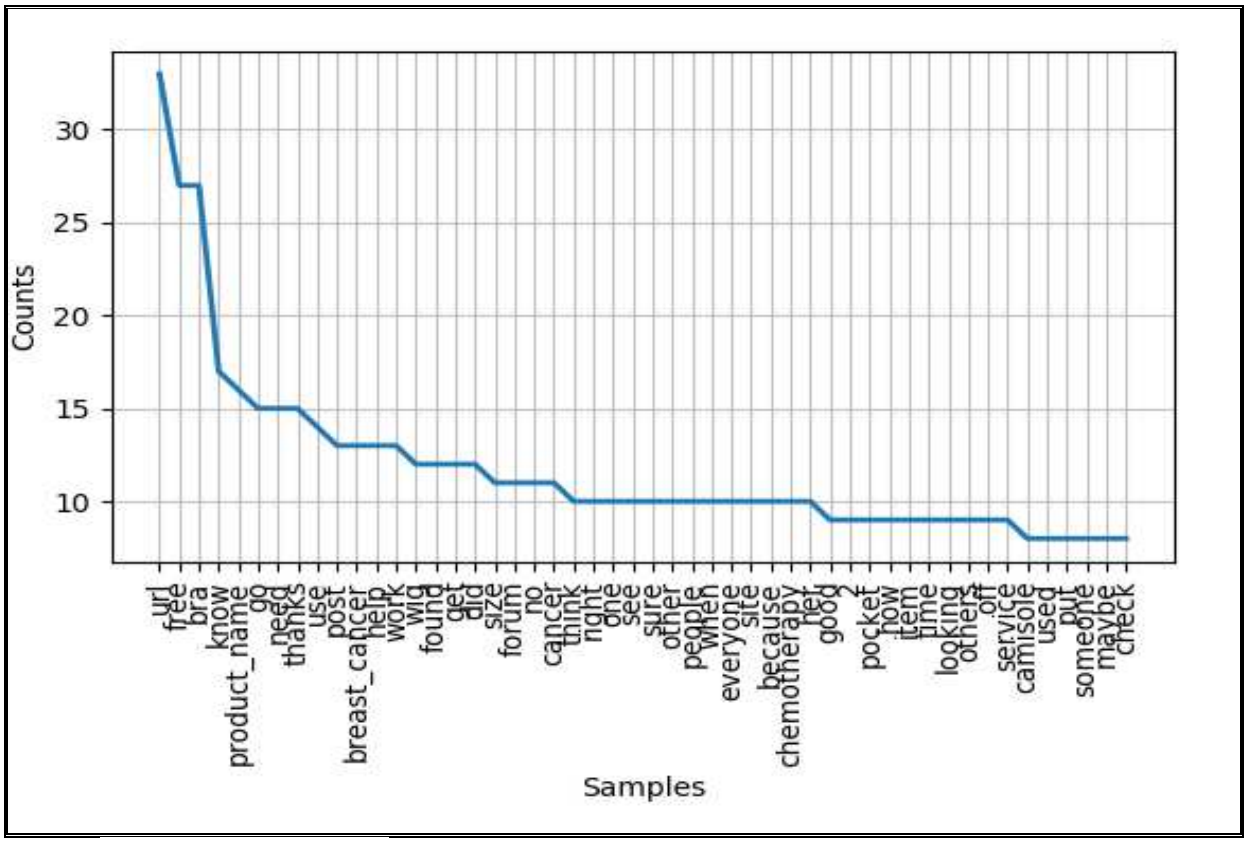


Figure 4. Top 50 tokens after normalizing and cleaning



4.2 Performance Evaluation of the Base Classifiers

Table 4 displays the results of four machine learning data models on manually labeled posts using three separate representations of the vectors with respect to precision, recall, F1-score, and accuracy. The F1-scores of the KNN, neural network, and SVM methods for each type of feature vector were less than 65% compared to the ENN, which had an F1 score of 72.1%. Among the encoding schemes, word2vec and doc2vec were significantly better than the TF-IDF representation. The reason could be, word2vec and doc2vec maintained the word context in the document, unlike TF-IDF, which did not find any word meaning and merely captured the words' frequencies throughout the document. Doc2vec also slightly outperformed word2vec, implying that the word order was

insignificant when considering the entire text. Finally, the ENN with the BERT-pre-trained vector was also compared with the doc2vec representation, as shown in Table 5. Interestingly, the performance of the ENN implemented with the BERT encoding showed slightly improved performance.

Table 4. Precision (P), Recall (R), and F1-scores of four data models over three vector representations

	P			**R**			**F1-score**			**Accuracy**		
	TF-IDF	**Word 2vec**	**Doc 2vec**	**TF-IDF**	**Word 2vec**	**Doc 2vec**	**TF-IDF**	**Word 2vec**	**Doc 2vec**	**TF-IDF**	**Word 2vec**	**Doc2vec**
ENN	0.698	0.723	**0.730**	0.719	0.729	**0.729**	0.702	0.725	**0.729**	**0.688**	**0.719**	**0.721**
Neural net-work	0.591	0.644	0.664	0.598	0.649	0.671	0.578	0.651	0.687	0.596	0.634	0.648
SVM	0.568	0.636	0.667	0.572	0.639	0.661	0.570	0.633	0.667	0.582	0.616	0.634
KNN	0.439	0.583	0.614	0.446	0.592	0.618	0.442	0.589	0.614	0.468	0.612	0.625

Table 5. Comparison of the Precision (P), Recall (R) and F1-scores of ENN-doc2vec and ENN-BERT

	P	**R**	**F1-score**
ENN-doc2vec	0.730	0.729	0.729
ENN-BERT	0.733	0.728	0.731

Figure 5. Confusion matrix for the predicted class in the ENN

N=200	**Predicted**						
Class 1 (ADE)	21	0	0	0	0	1	0
Class 2 (CLIN)	0	20	0	0	0	1	2
Class 3 (SUP)	0	0	33	4	2	1	0
Class 4 (CTXF)	3	0	3	30	1	0	
Class 5 (DXS)	0	0	2	0	22	0	0
Class 6 (FIN)	0	0	1	0	0	21	1
Class 7 (TXS)	0	0	0	2	0	0	29

The confusion matrix shown in Figure 5 also confirms the improved performance of the ENN, as shown by the nearly diagonal shape of the matrix. The false negative rate was higher for the "SUP" and "CTXF" classes.

4.3 Performance Comparison of Self-training classifiers

Table 6 shows the experimental results for the self-training classifier with the base classifier neural network (NN). The base classifier was implemented with the combined hypothesis H_{final} from the base classifier building phase. The self-training was performed with 40%, 50%, 60%, and 70% of the manually labeled posts. The testing was performed with 60%, 50%, 40%, and 30% of the (unseen) labeled posts. It is interesting to note that the self-training performance increased with the increase in the training size of the labeled posts. With k= 0.95 and 70% of the manually labeled training posts, self-training yielded an F1-score of 75%. In this way, 74,324 unlabeled posts were given labels, out of 100,000 unlabeled posts. The rest of the posts did not receive confident labels from the look-up list.

Table 6: Precision (P), recall (R) and F1-scores of self-training classifiers with different training data sizes

Classifiers	**Training size of the labelled data**											
	40%			**50%**			**60%**			**70%**		
	P	**R**	**F1-score**	**P**	**R**	**F1-score**	**P**	**R**	**F1-score**	**P**	**R**	**F1-score**
NN (Base classifier)	0.682	0.691	0.684	0.688	0.697	0.686	0.704	0.708	0.706	0.729	0.714	0.711
NN (k = 0.85)	0.684	0.694	0.687	0.682	0.693	0.684	0.708	0.711	0.709	0.730	0.726	0.727
NN (k = 0.90)	0.720	0.723	0.721	0.731	0.730	0.730	0.741	0.740	0.740	0.749	0.746	0.747
NN (k = 0.95)	0.729	0.728	0.728	0.738	0.736	0.770	0.746	0.745	0.745	0.751	0.749	**0.750**

4.4 Performance Comparison of Deep-learning Models

Table 7 shows the experimental results with three deep learning models. The models were tested with 20% of the manually labeled posts and 20% of the labeled posts from the self-training technique. All the models were implemented in python using the Keras package. The CNN was built with two convolution layers and three filters of size 3. The parameters were a dropout of 0.5, 10 epochs, and a batch size of 128. The LSTM was implemented with a batch size of 128, 256 LSTM units, and ten epochs with a learning rate of 0.01 and a dropout of 0.5. For the

BiLSTM, the same parameters for the LSTM model were repeated, but with 2 LSTM layers. As shown in the table, the LSTM achieved a significant improvement in accuracy of ~3%. However, the BiLSTM outperformed the LSTM with an increase in accuracy of ~1.5%. It is also noted that when the model was tested with semi-supervised labeled data, the performance was on par with the performance with the manually labeled test data. This result reiterates the effectiveness of the self-training technique with a look-up list in generating labels for the unlabeled data.

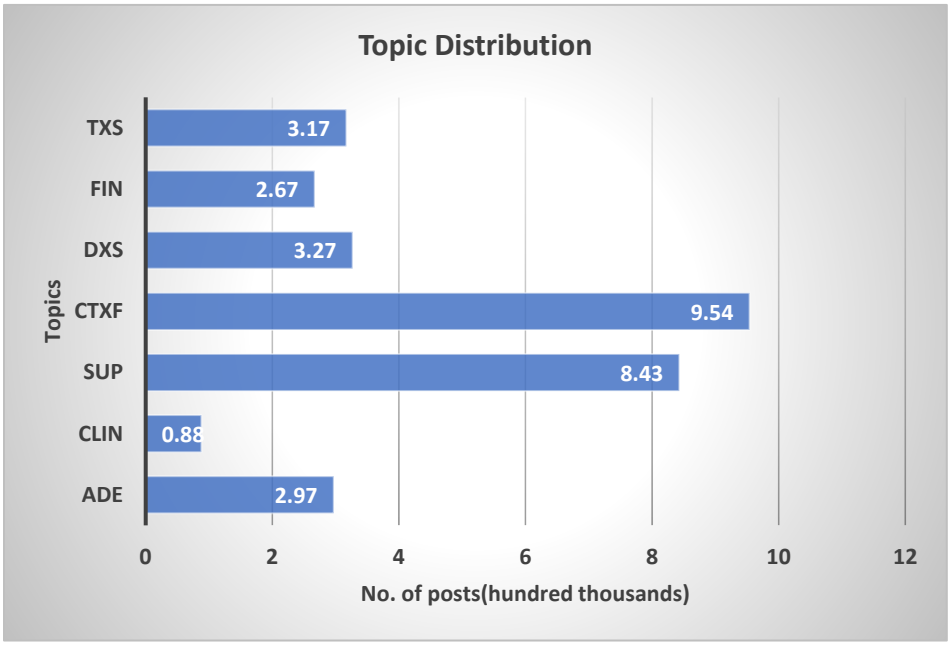
Table 7: Precision, recall and F1-scores of deep learning algorithms

Models	**20% of test data from semi-supervised labelled data**				**20% of test data from manually labelled data**			
	Precision	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
CNN	0.751	0.773	0.762	0.757	0.756	0.768	0.759	0.760
LSTM	0.780	0.789	0.784	0.782	0.779	0.791	0.789	0.786
BiLSTM	**0.792**	**0.790**	**0.791**	**0.797**	**0.795**	**0.789**	**0.793**	**0.795**

4.5 Overall Prevalence of Topics on entire Community Forums

To analyze the pervasiveness of topics over the breast cancer community forum, all the 3 million posts were classified by the best deep learning model, BiLSTM. The word embedding technique used was BERT. Posts with lengths greater than 500 words and less than ten words were eliminated. It was due to less information contained in those posts. Most lengthy posts conveyed personal stories regarding family, and short posts expressed gratitude, such as ‘thanks’ and ‘best wishes. Consequently, a few thousand posts were eliminated, and the rest of the posts were passed through the preprocessing and vectorization steps. Then, they were fed to the BiLSTM model for classification. The implementation was performed on an NVIDIA Tesla K80 GPU with 11 GB memory. The histogram of the topics is shown below in Figure 6. The classification reveals that the most discussed topics are ‘CTXF’ (contextual factors) and ‘SUP’ (support). It indicates that users more often discuss matters encountered during everyday life while going through the ‘cancer journey’ and informational, emotional, or material support provided by survivors or caregivers. The other topics, except CLIN (clinician or clinic), are at the same level. It has also been noted that financial matters are also a large concern for users, along with medical issues.

Figure 6: General prevalence of topics



5. Discussion

5.1 Principle Findings

This paper discusses the methods used to analyze OHC data for their potential impact on disease and wellness management. Clinicians tend to focus on a disease's clinical implications and may ignore patients' emotional well-being and daily lives. However, OHCs can be a source of support for patients. For several reasons, patients will probably not share many concerns with health care providers. It may be due to feeling that it is not necessary to discuss the issues due to embarrassment or not being conscious of the existence of a problem. This aspect is evident in the following conversation:

A post with the label SUP.REQ (Request for support): "*tell me about it I am so confused and I do not want to ask my dr it will just confuse me more. dcis was not even mentioned to me at the beginning, and yet it's there somehow, it's there the scans do not show everything. feeling very confused and fed up.*"

Answering post with the label SUP.OFF.IS (offering support in terms of information): "*dcis is often found with idc. they are not treating your dcis just your idc. It cannot respond to systemic treatments such as chemo, so the size of your dcis would not change from having chemo. sometimes after a biopsy or lumpectomy/mastectomy, they found more (or less) than they originally thought was there initially. I think a second opinion is a good idea in either case. wishing you the best whatever happens*!".

It would be impossible to find the "hidden" topics by analyzing only the forum names. The labels of topics in the current study provided an excellent pathway to obtain insights into what matters to patients other than clinical impact. The topics detected are more from the patients' perspective. For example, finding help from people in similar situations and their approaches to daily life is shown in the following post with the label SUP.Advice: *"i can offer some support. ….. your worth as a human being, and as a woman is not summed up in your breast. but your reaction is normal. I hope you can be good to yourself today and try to relieve some of the stress you are under. if it seems to call your md and tell them how anxious you are. You would not be the first person to do so, and many of us needed some anti-anxiety medication in the beginning. I believe you can get through this and I would hope the same for your relationship."*

The identified topics and labels are beneficial for health researchers. For instance, the data posts that come under the 'ADE' and 'TXS' can be further analyzed to determine the various drug-related issues while undergoing treatments. Similarly, the data posts under 'FIN' can unravel various insurance options that can be recommended to patients who are facing similar problems. Besides, negative or positive sentiment analysis towards a particular drug or any treatment options can be applied on these classified posts. These findings are an essential reference for both health researchers and clinical researchers to gain useful hypothesis.

Although the research in [14–15] explored the evolution of interactions on cancer forum communities and their satisfaction, identifying the issues raised using machine learning methods has not been sufficiently undertaken. The need to establish automated means of classifying topics in online forums to identify the posts that are crucial to community support has been noted by researchers [16–17]. As stated in Related Work, the similar study in [29] examined the breast cancer community's topics and achieved an overall F1-score of 65.4%. With their topic identification, two of our key topics, 'SUP' and 'FIN,' were not there. These two topics are an ideal direction for providing patients with personalized needs in terms of information or emotional support and financial suggestions/recommendations. Besides, with the limited number of labeled data and the data imbalance for each topic, the proposed BiLSTM model yielded a significant F1-score of 79.5%. This substantial improvement from the previous work could be attributed to the normalization and cleaning phase and the use of the BERT embedding

technique. The normalization and cleaning phase increased the text's lexical richness, and BERT successfully captured text semantics through contextual embedding techniques.

5.2 Limitations and Future Directions

The study was limited to only seven top-level labels. However, for comprehensive analysis, it is necessary to incorporate sublabels into the hierarchy of labels (as shown in Appendix 1) in the topic's name. The negligence of certain miscellaneous posts expressing emotions was one significant drawback of the study. Since emotions were conveyed in several posts, analyzing the posts from this viewpoint was essential to serve the patient's emotional needs. The base classifier's confusion matrix indicated that there were more missing values for "SUP" and "CTXF"; this will be investigated in future research. It is observed that some of the posts in these classes have both the labels "SUP" and "CTXF." It may be because users talk about both of these matters in the same posts. Therefore, these posts' multilabel nature, especially those belonging to these classes, must be further investigated. More in-depth research is required to explore the "support" the community members provide, specifically "support offers and support requests" and "expert survivors' recommendations," to satisfy the personal needs of patients.

Conclusion

This is a pioneering study on identifying the private concerns of patients through posts in breast cancer patient and survivor communities. Generally, the posts highlight diagnosis and treatment stories, adverse medication effects, information from various doctors, financial issues, family issues, and assistance rendered to patients by survivors, specifically materials, medication, and other vital information. The most active participants in the selected group are breast cancer survivors, who are willing to donate considerable numbers of treatment items. In this context, the emotional, informational, and material support provided by stories on social media in matters relating to diagnosis or treatment has gained great credence. Furthermore, the study demonstrates the effectiveness of software systems for evaluating online group discussions, which could have ramifications for other areas of human activities. The findings reveal the possibility of obtaining insights into patients' experiences in critical health management issues by studying social media.

Abbreviations

OHC: Online health community; NLP: Natural language processing; BERT: Bidirectional Encoder Representations from Transformers; ENN: Ensembled neural network; CNN: Convolutional neural network; LSTM: Long short-term memory; BiLSTM: Bidirectional long short-term memory.

Acknowledgements

The authors would like to acknowledge Breastcancer.org Community (website) as well as the graduate students from the BioHealth Informatics Department, IUPUI, who provided support in annotation part.

Authors' contributions

AB designed and implemented algorithms and classification models. AK and EZ collected dataset from Breastcancer.org, supervised the annotation process and reviewed the labels. JJ validated the labels. SM supervised the design and implementation of algorithms and classification models. AB conducted the experiments, JJ and SM validated the results. All the authors were involved in writing the manuscript and approved the final one.

Funding

This research received no specific grant from any funding agency in public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare that they have no competing interests.

Availability of data and material

The data that support the findings of this study are available from Breastcancer.org but restrictions apply to the availability of these data. Data are however available from the authors upon reasonable request and with permission of Breastcancer.org.

Ethics approval

Not applicable

Consent to participate

Not applicable

References

1. Kolowitz BJ, Lauro GR, Venturella J, Georgiev V, Barone M, Deible C, Shrestha R. Clinical social networking—a new revolution in provider communication and delivery of clinical information across providers of care?. Journal of digital imaging. 2014 Apr 1;27(2):192-9.

2. Medina EL, Mesquita CT, Loques Filho O. Healthcare social networks for patients with cardiovascular diseases and recommendation systems. Int J Cardiovasc Sci. 2016 Feb 13;29(1):80-5.

3. Miroševič Š, Prins JB, Selič P, Zaletel Kragelj L, Klemenc Ketiš Z. Prevalence and factors associated with unmet needs in post-treatment cancer survivors: A systematic review. European journal of cancer care. 2019 May;28(3):e13060.

4. Lo-Fo-Wong DN, de Haes HC, Aaronson NK, van Abbema DL, Admiraal JM, den Boer MD, van Hezewijk M, Immink M, Kaptein AA, Menke-Pluijmers MB, Russell NS. Health care use and remaining needs for support among women with breast cancer in the first 15 months after diagnosis: the role of the GP. Family practice. 2020 Feb;37(1):103-9..

5. Brandenbarg D, Maass SW, Geerse OP, Stegmann ME, Handberg C, Schroevers MJ, Duijts SF. A systematic review on the prevalence of symptoms of depression, anxiety and distress in long-term cancer survivors: Implications for primary care. European journal of cancer care. 2019 May;28(3):e13086.

6. Selove R, Foster M, Wujcik D, Sanderson M, Hull PC, Shen-Miller D, Wolff S, Friedman D. Psychosocial concerns and needs of cancer survivors treated at a comprehensive cancer center and a community safety net hospital. Supportive Care in Cancer. 2017 Mar 1;25(3):895-904.

7. Jones J, Pradhan M, Hosseini M, Kulanthaivel A, Hosseini M. Novel approach to cluster patient-generated data into actionable topics: case study of a web-based breast cancer forum. JMIR medical informatics. 2018;6(4): e45.

8. Nakikj D, Mamykina L. A park or a highway: Overcoming tensions in designing for socio-emotional and informational needs in online health communities. InProceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing 2017 Feb 25 (pp. 1304-1319).

9. Smith K, Golder S, Sarker A, Loke Y, O'Connor K, Gonzalez-Hernandez G. Methods to compare adverse events in twitter to faers, drug information databases, and systematic reviews: proof of concept with adalimumab. Drug safety. 2018 Dec 1;41(12):1397-410.

10. Hartzler AL, Taylor MN, Park A, Griffiths T, Backonja U, McDonald DW, Wahbeh S, Brown C, Pratt W. Leveraging cues from person-generated health data for peer matching in online communities. Journal of the American Medical Informatics Association. 2016 May 1;23(3):496-507.

11. Chomutare T, Årsand E, Fernandez-Luque L, Lauritzen J, Hartvigsen G. Inferring community structure in healthcare forums. Methods of information in medicine. 2013;52(02):160-7.

12. Wang X, Zhao K, Cha S, Amato MS, Cohn AM, Pearson JL, Papandonatos GD, Graham AL. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. Decision support systems. 2019 Jan 1;116:26-34.

13. Durant KT, McCray AT, Safran C. Modeling the temporal evolution of an online cancer forum. InProceedings of the 1st ACM International Health Informatics Symposium 2010 Nov 11 (pp. 356-365).

14. Vlahovic TA, Wang YC, Kraut RE, Levine JM. Support matching and satisfaction in an online breast cancer support community. InProceedings of the SIGCHI Conference on Human Factors in Computing Systems 2014 Apr 26 (pp. 1625-1634).

15. Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, Conway M. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. Journal of medical Internet research. 2017;19(2):e48.

16. Cabling ML, Turner JW, Hurtado-de-Mendoza A, Zhang Y, Jiang X, Drago F, Sheppard VB. Sentiment analysis of an online breast cancer support group: communicating about tamoxifen. Health communication. 2018 Sep 2;33(9):1158-65.

17. Elhadad N, Zhang S, Driscoll P, Brody S. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. InAMIA Annual Symposium Proceedings 2014 (Vol. 2014, p. 516). American Medical Informatics Association.

18. Yang, Christopher C., and Ling Jiang. "Enriching user experience in online health communities through thread recommendations and heterogeneous information network mining." IEEE Transactions on Computational Social Systems 5.4 (2018): 1049-1060.

19. Liu Y, Xu S, Yoon HJ, Tourassi G. Extracting patient demographics and personal medical information from online health forums. InAMIA Annual Symposium Proceedings 2014 (Vol. 2014, p. 1825). American Medical Informatics Association.

20. Nguyen LH, Salopek A, Zhao L, Jin F. A natural language normalization approach to enhance social media text reasoning. In2017 IEEE International Conference on Big Data (Big Data) 2017 Dec 11 (pp. 2019-2026). IEEE.

21. Lee K, Hasan SA, Farri O, Choudhary A, Agrawal A. Medical concept normalization for online user-generated texts. In2017 IEEE International Conference on Healthcare Informatics (ICHI) 2017 Aug 23 (pp. 462-469). IEEE.

22. Clark E, Araki K. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences. 2011 Jan 1;27:2-11.

23. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearbook of medical informatics. 2019 Aug;28(1):208.

24. Momtazi S, Rahbar A, Salami D, Khanijazani I. A Joint Semantic Vector Representation Model for Text Clustering and Classification. Journal of AI and Data Mining. 2019 Jul 1;7(3):443-50.

25. Singh AK, Shashi M. Vectorization of Text Documents for Identifying Unifiable News Articles. Int. J. Adv. Comput. Sci. Appl. 2019;10.

26. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics: X. 2019 Dec 1;4:100057.

27. Chen, Langtao. "A Classification Framework for Online Social Support Using Deep Learning." International Conference on Human-Computer Interaction. Springer, Cham, 2019.

28. Zhu, Binjun, Xiaofeng Cai, and Ruichu Cai. "Answer Quality Evaluation in Online Health Care Community." 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). Atlantis Press, 2018.

29. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. Journal of biomedical informatics. 2017 May 1; 69:1-9.

30. Keyvanpour MR, Imani MB. Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms. Intelligent Data Analysis. 2013 Jan 1;17(3):367-85.

31. Sigdel, Madhav, et al. "Evaluation of semi-supervised learning for classification of protein crystallization imagery." IEEE SOUTHEASTCON 2014. IEEE, 2014.

32. Gowda HS, Suhil M, Guru DS, Raju LN. Semi-supervised text categorization using recursive K-means clustering. InInternational Conference on Recent Trends in Image Processing and Pattern Recognition 2016 Dec 16 (pp. 217-227). Springer, Singapore.

33. Seeger M. Learning with labeled and unlabeled data. technical report, University of Edinburgh, Tech. Rep. 2001

34. Jalan R, Gupta M, Varma V. Medical forum question classification using deep learning. In European Conference on Information Retrieval 2018 Mar 26 (pp. 45-58). Springer, Cham.

35. Mayring P. Qualitative content analysis forum qualitative sozialforschung. InForum: qualitative social research 2000 Jun (Vol. 1, No. 2, pp. 2-00).

36. Holden RJ, Kulanthaivel A, Purkayastha S, Goggins KM, Kripalani S. Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. International journal of medical informatics. 2017 Dec 1; 108:158-67.

37. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.

38. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).

39. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

40. Chen Y, Chang H, Meng J, Zhang D. Ensemble Neural Networks (ENN): A gradient-free stochastic method. Neural Networks. 2019 Feb 1; 110:170-85.

41. Zhou ZH, Wu J, Tang W. Ensembling neural networks: many could be better than all. Artificial intelligence. 2002 May 1;137(1-2):239-63.

42. Polikar R, Upda L, Upda SS, Honavar V. Learn++: An incremental learning algorithm for supervised neural networks. IEEE transactions on systems, man, and cybernetics, part C (applications and reviews). 2001 Nov;31(4):497-508.

Figures

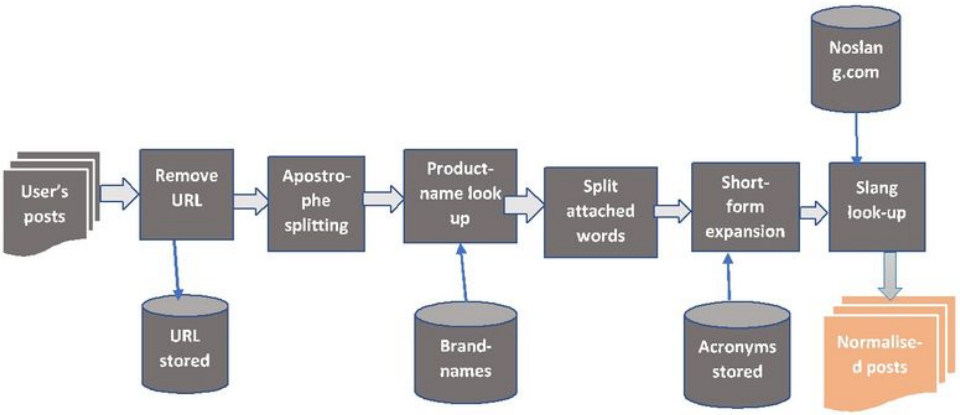


Figure 1

Framework for Normalizing and Cleaning

Input: 1) Data set D with n samples $((X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n))$ where $X_1, X_2 \ldots X_n$ are n data points in the data set and $Y_1, Y_2 \ldots Y_n$ are the n corresponding annotated labels.

2) $W_i = 1/n$ where the initial weight for each data points, $i=1$ *to* n

3) A classifier

4) 'm' number of iterations

Output: Combined hypothesis, $H_{final} = \sum_1^m h_j(Score_j)$

while $j=1, 2 \ldots m$ or classifier's error stabilize

1) Randomly choose TR_j and TE_j from data set D. Create hypothesis h_j: $X \rightarrow Y$ on $TR_j + TE_j$
2) Call weak classifier, provided with TR_j and Calculate the classification error on TE_j, $\varepsilon_j = \sum_{i:h_j(x_i \neq y_i)} W_i(j)$ on TE_j
3) If $\varepsilon_j > \frac{1}{2}$, discard h_j and go to step 1, otherwise, normalize error, $\beta_j = \varepsilon_j / (1 - \varepsilon_j)$
4) Calculate $score_j = \sum_{i:h_j(x_i=y_i)} \log\left(1/\beta_j\right)$, for $\forall$ Y
5) Update weight: $W_i(j+1) = W_i(j) \times \begin{cases} 1, \text{if } h_j(x_i \neq y_i) \\ \beta_j, \text{otherwise} \end{cases}$

$$Combined\ Score_j = \underbrace{\operatorname{argmax}}_{y \in Y} score_j$$

Figure 2

Algorithm for the ENN

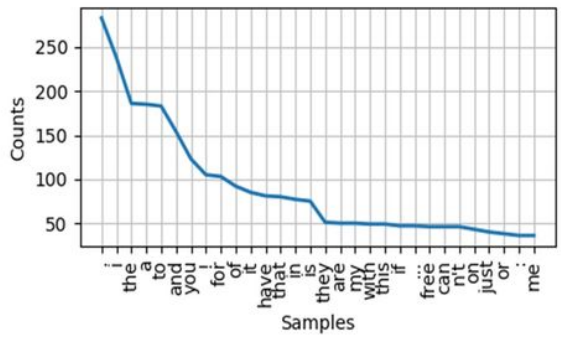


Figure 3

Top 50 tokens before normalizing and cleaning

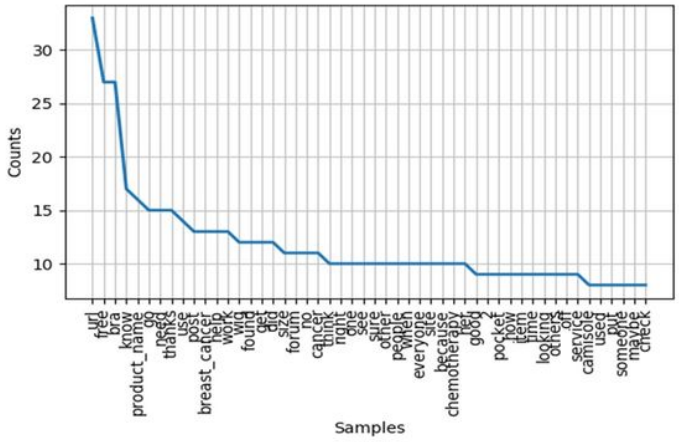


Figure 4

Top 50 tokens after normalizing and cleaning

N=200	Predicted						
Class 1 (ADE)	21	0	0	0	0	1	0
Class 2 (CLIN)	0	20	0	0	0	1	2
Class 3 (SUP)	0	0	33	4	2	1	0
Class 4 (CTXF)	3	0	3	30	1	0	
Class 5 (DXS)	0	0	2	0	22	0	0
Class 6 (FIN)	0	0	1	0	0	21	1
Class 7 (TXS)	0	0	0	2	0	0	29

Figure 5

Confusion matrix for the predicted class in the ENN

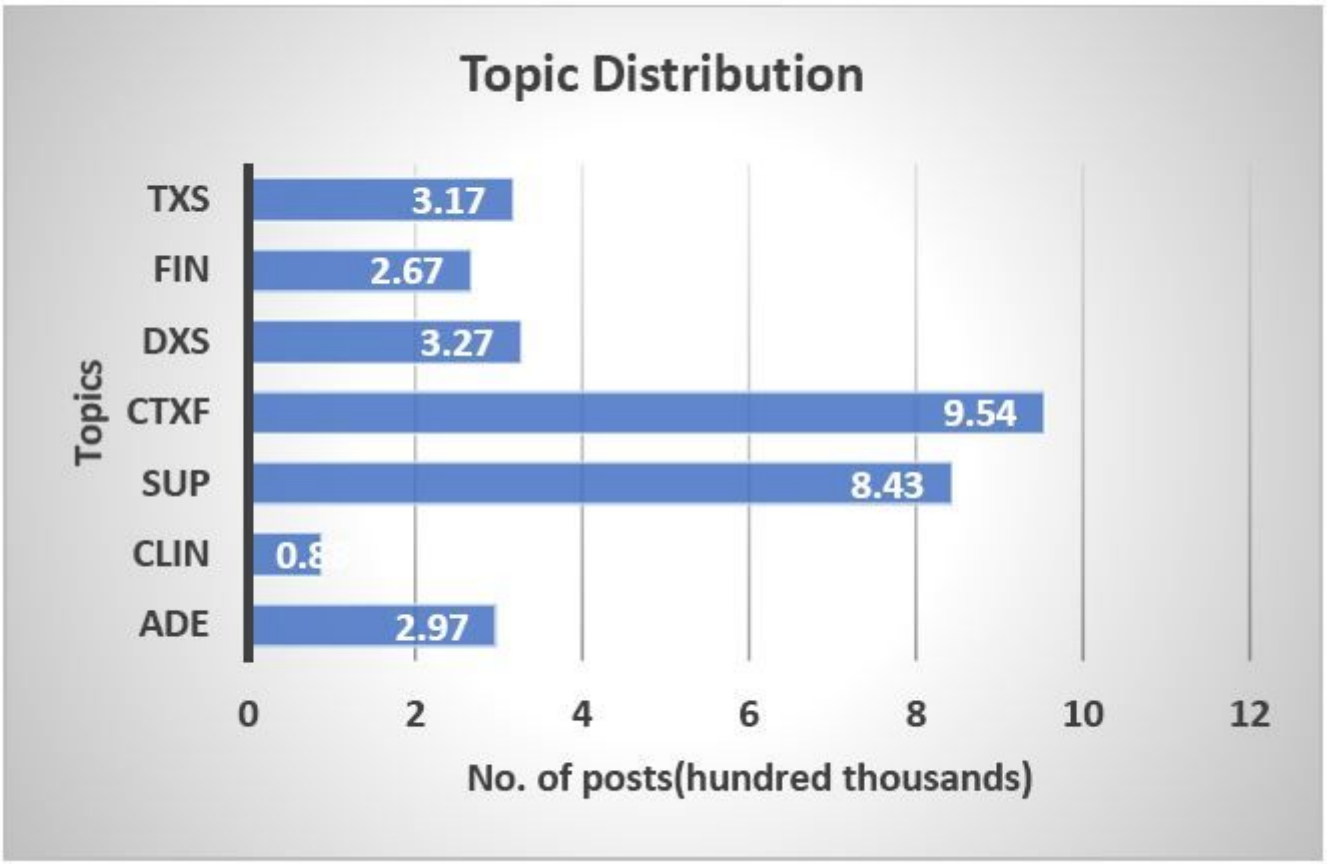


Figure 6

General prevalence of topics

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1.docx
- Appendix2.docx