Spring 2018

Trajectory-based methods to predict user churn in online health communities

Apoorva Joshi
University of Iowa

TRAJECTORY-BASED METHODS TO PREDICT USER CHURN IN ONLINE HEALTH COMMUNITIES

by

Apoorva Joshi

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Electrical and Computer Engineering in the
Graduate College of
The University of Iowa

May 2018

Thesis Supervisors: Assistant Professor Kang Zhao
Professor Sudhakar M. Reddy

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Apoorva Joshi

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Electrical and Computer Engineering at the May 2018 graduation.

Thesis Committee: _____
Kang Zhao, Thesis Supervisor

_____
Sudhakar M. Reddy

_____
Xun Zhou

_____
Guadalupe M. Canahuate

ACKNOWLEDGEMENTS

At the outset, I would like to express my sincere thanks to Prof. Kang Zhao, my thesis supervisor for his guidance and encouragement throughout the course of my master's thesis at The University of Iowa. He has been very inspiring and motivating and, despite his exceedingly busy schedule, ensured that he was always available for me to discuss any project-related problems even at short notice. I have been extremely fortunate to get an opportunity to be a part of his research group for the past two semesters. I am also extremely grateful to Prof. Sudhakar M. Reddy, my academic advisor for periodically reviewing my progress, providing useful suggestions and ensuring that all administrative issues were always taken care of.

I would also like to thank Dr. Xun Zhou for all his inputs throughout the course of the thesis project. Additionally, I would also like to thank everyone in Prof. Zhao's research group. The regular discussions that I had with the group members during the weekly project meetings further helped me considerably broaden my knowledge base.

Last but not the least; I would like to thank my family and friends, for their love, endless support, and encouragement.

ABSTRACT

Online Health Communities (OHCs) have positively disrupted the modern global healthcare system as patients and caregivers are interacting online with similar peers to improve quality of their life. Social support is the pillar of OHCs and, hence, analyzing the different types of social support activities contributes to a better understanding and prediction of future user engagement in OHCs.

This thesis used data from a popular OHC, called Breastcancer.org, to first classify user posts in the community into the different categories of social support using Word2Vec for language processing and six different classifiers were explored, resulting in the conclusion that Random Forest was the best approach for classification of the user posts. This exercise helped identify the different types of social support activities that users participate in and also detect the most common type of social support activity among users in the community.

Thereafter, three trajectory-based methods were proposed and implemented to predict user churn (attrition) from the OHC. Comparison of the proposed trajectory-based methods with two non-trajectory-based benchmark methods helped establish that user trajectories, which represent the month-to-month change in the type of social support activity of users are effective pointers for user churn from the community.

The results and findings from this thesis could help OHC managers better understand the needs of users in the community and take necessary steps to improve user retention and community management.

PUBLIC ABSTRACT

Breast cancer is the most common cause of mortality in women. 1 in 8 women in the United States develop invasive cancer in the course of their lifetime. Online Health Communities (OHCs) are a great source of comfort, knowledge and support for patients and their caregivers. Interacting online with similar peers to gain comfort, knowledge and support helps improve quality of life. Social support is the pillar of OHCs and can be classified into three main categories, namely companionship, emotional support and informational support.

This thesis project first aimed at classifying user activity into different categories of social support. Thereafter, trajectory-based techniques were used to analyze how a change in the social support activity of a user over time affected user attrition from the community.

The results and key findings made in this thesis can enable better management and sustenance of successful OHCs. Detecting different types of social support activities can help OHC administrators better understand the trends in user participation in the community. Analyzing patterns in user churn can help facilitate targeted interventions, through alerts and recommendations for user retention.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

OHC	Online Health Community
SVM	Support Vector Machine
BOW	Bag of Words
CBOW	Continuous Bag of Words
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
AUC	Area Under Curve
COM	Companionship
PES	Providing Emotional Support
PIS	Providing Informational Support
SES	Seeking Emotional Support
SIS	Seeking Informational Support
TFIDF	Term Frequency- Inverse Document Frequency
Oth	Other

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

The global healthcare system is undergoing a radical shift as more and more patients use technology to take control of their personal well-being. A field once dominated by academia and leaders in the pharmaceutical industry now has a third category of stakeholders, who have the potential to change the dynamics of the healthcare industry in ways not imagined before. This group comprises patients, caregivers, friends, family members and interested citizen-scientists who rally around various healthcare conditions and concerns and bond over their experiences, frustrations and questions on the Internet [1-2].

According to a study by the Pew Research Center, 80% of adult Internet users in the United States use the Internet for health-related purposes. Among them, 34% read health-related experiences or comments from others. While the use of social media tools and online health communities (OHCs) as hubs to achieve breakthroughs in medical innovation may appear far-fetched, their importance cannot be overstated. However, as with any participant-driven platform, online health communities run a potential risk of falsified data, misinterpretation of terminologies and sporadic participation of users. Nonetheless, it is abundantly obvious that the above online communities have made communication multi-directional, information more easily accessible and problem sharing / solving simple and fast. This has resulted in patients becoming more engaged in their healthcare via technologies accessible to them. In an industry that spends trillions of dollars a year, it can pay both figuratively and otherwise, to harness the immense potential of OHCs [3-6].

In view of their growing relevance in present-day patient care, observational research of user behavior in OHCs could be valuable in several ways. The outcome of such studies can enable better management and sustenance of successful OHCs, which can have a tremendous impact on patient care as well as contribute towards improving the quality of

health research and innovation. The success of an OHC can be measured by the extent of user participation and user churn (i.e., users leaving the community). A successful OHC is one which has robust user participation and relationship building between members. Poor participation and transient membership leads to the 'failure' of an online community and its eventual irrelevance.

The benefits of continued and active participation in an OHC are two-fold. On one hand, receiving support from the community is empowering and helps patients deal with the stress of coping with their medical condition [7-9]. On the other hand, providing support to others in the community can be therapeutic and satisfying [9]. There have been scientifically proven relationships between posting frequency and psycho-social well-being [10]. There is a possibility that individuals who have already received the anticipated support from an OHC, or recovered from the disease, may consider leaving the OHC. This is clearly undesirable as it adversely impacts the utility of the OHC. Even though user-provided information about a disease will still be available on the Internet to new OHC members, individual users can be deprived of significant psycho-social benefits if there is an exodus of experienced users from the OHC [10]. In fact, assistance for new members from experienced members and continuous participation of as many members as possible are also key factors for the success of online communities [5]. Therefore, better understanding and accurate prediction of users' participation in OHCs can help to build and sustain a successful OHC through improved community design, management, and user retention [2].

As discussed above, seeking, receiving and providing different kinds of social support are the very crux of OHCs. Despite extensive research in the realm of OHCs, very few studies have explored the relationship between the different types of social support and participation of users in the OHC. This project uses language processing, machine learning, trajectory clustering and predictive modeling techniques on large-scale data from a real-world OHC to investigate user involvement in the OHC based on the type of social support activities the user participates in. Additionally, it also aims at predicting future user engagement in the community based on past activity.

1.2 BACKGROUND

According to Shumaker and Brownell [11], social support refers to the "exchange of resources between at least two individuals perceived by the provider or the recipient to be intended to enhance the well-being of the recipient." Based on the nature of exchanged "resources," community psychology researchers have identified different types of social support [11-13]. This thesis broadly deals with four kinds of social support identified [14-15], namely emotional support, informational support, companionship and instrumental support. These are defined as follows:

Emotional support: Involves expression of concern, comfort, sympathy, empathy, encouragement, affection and understanding among users.

Informational support: Usually deals with advice, information, personal experience regarding the medical condition, discussions regarding medication, hospitals and other technicalities surrounding healthcare.

Companionship: Also known as network support, this comprises informal chatting, humor, teasing and discussion of daily offline activities not necessarily related to one's medical condition.

Instrumental support: Refers to a more physical aspect of providing support offline. Such support could range from helping with transportation to the hospital in case of an emergency to offering to help peers with grocery shopping.

Empirical studies have shown that emotional, informational support and companionship are common in OHCs while instrumental support is rare as geographical constraints limit it. Also, instrumental support often occurs through private messaging platforms like email, text messaging etc., instead of a public forum. Hence, instrumental support has not been considered for the purpose of this thesis.

The emergence of OHCs as a key "stakeholder" category has radically enhanced the scope of online social support research. Traditional studies of offline support communities conducted through questionnaires, surveys and interviews of community

members face several challenges, particularly concerning data collection. Firstly, the scale of the data collected is limited by the time and labour involved in conducting surveys, interviews etc. Secondly, the scope of information obtained in such studies could be skewed based on the sampling pool for the study. For example, if only those members who had positive outcomes from their medical treatment participated in the survey, the outcome and conclusions of a research study performed on this sample pool would be largely biased. The largest drawback of such methods is that they lack temporal granularity. It is very difficult to accurately track the activities of members across an extended time period. Even if the study is conducted periodically, say every few months, considerable crucial information is lost in the time between studies.

In the above context, OHCs not only enable but also record asynchronous and distributed web interactions among individuals. This makes large amounts of data available for computational analysis, thus solving most of the major problems associated with manually administered studies. The data so recorded has valuable information about the number, type and time associated with online user interactions, thereby substantially increasing the granularity of research studies performed on the data.

1.3 GOALS OF THE THESIS

In order to analyze social support activity on a large scale, it is important to have an automated method to reveal the type of social support embedded in user contributions in the community. This provides the motivation for the goals targeted by this thesis.

Goal 1: Mine large-scale unstructured text data contributed by OHC users to detect the type of social activities of the users in the community.

Goal 2: Explore trajectory-based methods to predict user churn from the community, based on change in user activity in the community over time.

1.4 OUTLINE OF THE THESIS

Apart from this introductory chapter providing an overview and background of the project work undertaken and identifying specific research goals targeted to be accomplished during this thesis, the rest of this thesis is organized as follows:

Chapter 2 provides a brief overview of machine learning, description of terminologies associated with classification and also the different evaluation metrics used to validate the classification algorithms used in the thesis. It also gives an outline of previous research work in the context churn prediction.

Chapter 3 provides a comprehensive description of tools, frameworks and methods that were used to carry out the first task of the project i.e. classification of social support, followed by a detailed discussion of results and outcomes of the classification.

Chapter 4 explains the implementation of three trajectory-based methods for churn prediction and two non-trajectory-based benchmark methods. It also consists of a compilation of key outcomes and findings obtained as a result of a comparative study across the trajectory-based and benchmark methods.

Finally, Chapter 5 includes a summary of the research carried out and the scope for future study in this field of immense relevance.

CHAPTER 2

THEORY

2.1 MACHINE LEARNING

Machine learning tasks can be divided into two main categories, namely Supervised Learning and Unsupervised Learning. Supervised Learning refers to the category where the algorithm has input, and output variables and the goal is to approximate a mapping function based on already available data which can then predict the desired output for new input data. In unsupervised learning, the output of the learning task is not well-defined, and the goal is to find some structure or patterns in the input data. This thesis deals with classification, where the target labels are known, thus making it a supervised learning task.

This thesis utilizes various simple classification algorithms, namely Naïve Bayes, Logistic Regression, Support Vector Machines, Decision Tree and complex ensemble classification algorithms like Random Forest and Adaboost.

2.2 LINEAR CLASSIFICATION

Given a set of data points, each belonging to one of two classes, the goal of classification is to determine the class of a new data point based on the existing ("training" in Machine Learning terminology) dataset. A linear classifier makes this decision based on the value of a linear combination of the feature vector. The feature vector is a combination of "features", which are quantitative or qualitative translations of the observations in the training dataset. For a classification task, the feature vectors of the training set are fed to the algorithm, along with their corresponding class, thus making it a supervised learning problem. This thesis project explored various simple classification algorithms, namely Naïve Bayes, Logistic Regression and Support Vector Machines.

2.3 ENSEMBLE CLASSIFICATION

Ensemble classifiers, as the name suggests, pool the predictions of multiple simple base models. There is enough empirical and theoretical evidence to show that ensemble

classification facilitates an increase in the prediction accuracy. There are two types of algorithms for ensemble models- bagging algorithms, which derive independent base estimators from bootstrapped samples of original data and boosting algorithms which iteratively add weak estimators that are trained to avoid errors of the current ensemble. This thesis explores a standalone Decision tree algorithm, Random Forest, a popular bagging algorithm, in combination with Decision Tree, and Adaboost, which is a boosting algorithm.

2.4 CLASSIFICATION EVALUATION METRICS

Several evaluation metrics were calculated and analyzed before deciding the best classification model for this thesis. A brief description of each of the metrics is as follows [16]:

2.4.1 K-fold Cross Validation

Cross validation is a model validation technique used to assess the extent of generalization of a model over an independent dataset. Cross validation is an important metric in prediction problems like classification in order to evaluate the accuracy of the prediction. In k-fold cross validation, the original data is randomly shuffled and divided into k equal-sized batches. Out of the k batches, one batch is retained as the validation set and the remaining k-1 batches are used as the training set. This process is repeated k times, with each of the k samples used for validation exactly once. The prediction results from the k folds are averaged to arrive at a single estimation.

2.4.2 Accuracy and Other Metrics

The accuracy of a classification model is simply the number of correct class predictions made over the total number of predictions. However, a model cannot be labeled as "good" unless it is robust. There are several parameters that provide information about the robustness of a model. These parameters make use of certain terminologies, namely True Positives (TP) comprising correctly identified data points, False Positives (FP) comprising incorrectly identified data points, True Negatives (TN) consisting of correctly

rejected data points and False Negatives (FN) consisting of incorrectly rejected data points.

The metrics used in this thesis are as below:

- Sensitivity/ Recall: Also known as the True Positive Rate is the ratio of True Positives to the sum total of True Positives and False Negatives.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad \textbf{(2.1)}$$

- Specificity: Also known as True Negative Rate is the ratio of the true Negatives to the sum total of True Negatives and False Positives.

$$Specificity = \frac{TN}{(TN+FP)} \quad \textbf{(2.2)}$$

- Precision: Also known as True Positive Value is calculated as the ratio between True Positives and the sum total of True Positives and False Positives.

$$Precision = \frac{TP}{(TP+FP)} \quad \textbf{(2.3)}$$

- F1 Score: Considers both the Precision and Recall of an experiment to measure its accuracy. Traditionally, F1 Score is the harmonic mean of Precision and Recall.
- ROC-AUC Score: An ROC curve plots the Sensitivity (y-axis) of a test against Specificity(x-axis). The AUC, which is the area under the ROC Curve is a measure of the accuracy of the test. An area of 1 is representative of a perfect test while an area of 0.5 represents a failed test. Figure 2.2 illustrates the concept of AUC.

Accuracy and AUC Score were both used to evaluate the results of the classification task in the project. Precision and Recall were focused upon, in addition to accuracy and AUC to assess the results of the churn prediction model built thereafter.

2.5 PREVIOUS WORK ON CHURN PREDICTION

Churn prediction has been a topic of major research interest in online communities and forums, given that the success of these platforms is based on their user retention and following. Several avenues have been explored to predict users as future churners or non-churners based on historical data.

Lemmens and Croux have used traditional statistical methods such as logistic regression to predict churn [17]. Verbeke et al. have compared techniques like rule-based classifiers (Ripper, PART), decision tree approaches (C4.5, CART and Alternate Decision Trees) and Ensemble Methods (Random Forest, Logistic Model Tree, Bagging and Boosting) to provide insights for churn prediction in the telecommunication sector [18]. Datta et al. have used non-statistical models like K-Nearest Neighbor for churn prediction in the automated modelling space [19]. Mozer et al. have explored neural networks, to find that they perform better than models like logistic regression for churn prediction [20].

Most of the above studies have evaluated traditional classification models on a single churn prediction dataset and hence, development of more sophisticated models, which are specific only to churn prediction remains an open research issue. This thesis proposes a combination of trajectory mining techniques with pre-existing classification methods, to predict churn.

CHAPTER 3

CLASSIFICATION OF SOCIAL SUPPORT

This chapter provides a detailed description and discussion of all the methods used for the Natural Language Processing and Text Classification tasks of the thesis project. It also gives an overview of the tools and frameworks used to perform the above tasks, along with a comprehensive explanation of the scale and nature of the data used in the project. The results and key findings from the above studies have also been attached towards the end of this chapter.

3.1 THE DATASET AND TAXONOMY OF SOCIAL SUPPORT

The dataset for this thesis consists of user posts taken from Breastcancer.org, a popular peer-to-peer online health community for breast cancer survivors and their caregivers.

It comprises all public posts available on Breastcancer.org between October 2002 and August 2013. There are more than 2.8 million user posts, including 107549 initial posts (main threads, not including comments), contributed by around 50,000 different users. Each entry in the dataset consists of three fields namely User ID, the user post and the Timestamp of the post.

As discussed previously, barring instrumental support, there are five main categories of social support in OHCs- Companionship(COM), Providing Emotional Support(PES), Providing Informational Support(PIS), Seeking Emotional Support(SES) and Seeking Informational Support(SIS). 1333 randomly selected, pre-annotated posts from this dataset were used to train various classification algorithms in order to classify the rest of the data into different categories of social support. Five human annotators were trained on the definitions and examples of each of the social support categories. They then read each of the 1333 selected posts and decided one or more categories for each post. A pool of 10 posts annotated by domain experts were also added to the pool, for quality control. For each post, the results from only those annotators who featured in the top three for the 10 quality control posts were retained and the rest discarded. A majority vote among the top

three annotators was finally used to decide whether or not a post belonged to a particular category [2].

3.2 TOOLS

The thesis uses several tools for Natural Language Processing and Machine Learning. Python (Version 2.7) was the coding language of choice, given its plethora of tools, frameworks and libraries that could be leveraged to perform several tasks during the thesis project. The classification models in Scikit-Learn have been used extensively in this thesis. Tools like Numpy, Pandas and Pickle were used for ease of loading and storing large amounts of data, data cleaning and pre-processing. The NLTK and Gensim toolkits were used for the language processing task of the thesis [21-22]. Seaborn, Matplotlib and Plotly, Python's visualization libraries, were used extensively for data analysis throughout the thesis. An instance of the Google Compute Engine, consisting of 4 CPUs and 1 GPU was used to run several classification jobs, given the large scale of the datasets used in the project [23].

3.3 DATA PRE-PROCESSING

The biggest challenge with text data today is that most of it is unstructured i.e. it is either present in data silos or scattered across many digital archives. In order to produce any actionable insights from this data, it is important to convert the raw, unstructured data into a cleaner form. Social media data, in particular, is highly unstructured since user posts can consist of typographical errors, incorrect grammar, slang words or even unwanted content like URLs, stop words and other stray characters. Therefore, the first step of this thesis project was to clean and prepare the user posts in the dataset into a form compatible with the classification models to be used. The different stages in the cleaning and preparation step are detailed below:

1) Decoding the data: This is the process of converting the complex symbols and characters in the data into a form understood by the computer. The data maybe subject to different types of decoding. 'Latin1', a widely accepted format for text data, was used in this thesis.

2) Case Desensitization and Tokenization: The case of the words is inconsequential while converting text into vectors. Hence, the entire user post was converted to lowercase before tokenization. Tokenizing a string of text splits the text based on a delimiter. The user posts in our data set were tokenized based on the space delimiter to split the sentences in the posts into words.
3) Removal of Stop words: Stop words are commonly occurring words which have little contribution to the context or semantics related to the domain of the document, for example, articles (a, an, the), is, at, which, on etc. Removing stop words does little harm and helps in saving computational memory as well as time. The NLTK module in Python, which consists of a list of common stop words, was used as reference to remove stop words from the data.
4) Removal of punctuation and stray characters: Regular expressions are a useful way of extracting specific content in text data. Punctuations and other arbitrary characters in the user posts were removed by writing a regular expression that extracts only word characters from the words.

As a result of the tokenization and string cleaning processes, the user posts were converted into a list of case-insensitive words, devoid of stop words, punctuation and strays.

For the 1333 data entries to be used for training the classification model, the final step in the pre-processing stage was to convert the categorical social support type labels into binary labels. The general convention of '1' if the post belongs to the said category of social support and '0' if it does not belong to the category, was used. This process was repeated for all the five categories of social support.

3.4 NATURAL LANGUAGE PROCESSING MODELS

Prior to building the text classification model on the user posts, it was important to decide and build the features or inputs to the model. The most common baseline approach for text classification problems is the Bag of Words (BOW) model. This thesis used a more sophisticated model called Word2Vec, which goes slightly beyond the capacities of the BOW model. A brief overview of both the models is as follows:

3.4.1 Bag of Words (BOW)

Bag of Words is a vector representation describing the occurrence of words in the text, given a vocabulary of known words. It is called a "bag" of words as it disregards the relative position or structure of the words and is only concerned with the occurrence of the known word in the document. The simplest method to build the word vector representation is to assign binary values - i.e., '1' for presence and '0' for absence of the word in the document. To illustrate the intuition behind Bag of Words, consider a dataset consisting of two sentences:

Sentence 1: The fox jumped over the lazy dog

Sentence 2: The girl jumped over the wall

Table 3.1 represents the BOW representation of the two sentences

Vocabulary	**The**	**Fox**	**jumped**	**over**	**lazy**	**Dog**	**girl**	**wall**
Sentence 1	1	1	1	1	1	1	0	0
Sentence 2	1	0	1	1	0	0	1	1

Table 3.1: Illustration of BOW

Due the simplicity of the BOW model, it has several limitations. Firstly, the vocabulary must be designed carefully in order to manage memory space efficiently. Secondly, binary vector representations are sparse, and hence harder to model, both for computational reasons and because they contain very little information in a large representational space. Lastly, disregarding the pattern and position of words in the document results in loss of context and semantics, which can potentially contribute a lot to the model.

3.4.2 Word2Vec [24]

Word2Vec is a word embedding technique which captures the degree of similarity between words. In natural language processing, word embedding is a set of feature

learning techniques which map words into vectors of real numbers, given a vocabulary of known words. Word2Vec produces these vectors such that words with similar context occur close to each other in the vector space, thus contributing some semantic information to the model. Word2Vec has two different implementations, namely Continuous Bag of Words(CBOW) and Skip-gram, which differ slightly in their characteristics. CBOW considers a window of words around the current word and can be viewed as prediction of a word, given the context. Similar to the traditional BOW, CBOW also does not consider the order of the words. Skip-gram does quite the opposite, wherein it predicts the context, given a word. It does so by assigning lower weights to distant or "out of context" words. Although Skip-gram demands higher computational time, the quality of the word vectors produced using Skip-gram is high, even for infrequently occurring words and it adds semantic information to the model. The text classification task involved in this thesis involves prediction of the type of social support based on the text. For this purpose, it is important for the model to understand the context and semantics behind the text, thus providing the rationale behind using the Skip-gram Word2Vec model in this thesis.

3.5 FEATURE EMBEDDING

The Gensim library in Python was used to build a Word2Vec model to extract features for classification [22]. Word2Vec results in word vectors that carry some semantic information about the words, which is not captured by other language processing models.

The inputs to the Gensim model were a space-padded version of the word lists obtained after tokenizing and string cleaning i.e. blank spaces were padded to the end of "shorter" lists to make the lengths of all word lists equal to that of the longest word list to maintain dimensional consistency of the word lists.

Gensim creates word embeddings from the word lists. Embeddings, as the name suggests, are vectors of real numbers, positioned such that they carry "embedded" information about the context of the text. The model used in this thesis performs two operations on the word lists, namely Term Frequency- Inverse Document Frequency (TFIDF) Vectorization and Dimensionality Reduction. TFIDF goes beyond one- hot encoding of

words and weights the words based on their frequency of occurrence in the text, resulting in weighted real-valued word vectors. Dimensionality Reduction is done to retain only the principal, maximally informative dimensions in the vector space. In this thesis, vector representations of length 70,000 were compressed into word embeddings of length 300.

3.6 CLASSIFICATION

Six different machine learning classifiers were compared and analyzed to decide on a final model for classification of the user posts into their respective categories of social support. Five classifiers were built for each of the social support categories. Three simple classifiers, namely Naïve Bayes, Logistic Regression and Support Vector Machine and three ensemble classifiers, namely Decision Tree, Random Forest and Adaboost were evaluated using 10- Fold cross-validation, with focus on the ROC-AUC metric to decide the best performing classifier.

The pre-annotated dataset of 1333 user posts was used for training and testing the classifiers. The word embeddings of these posts were used as the features for the models.

The steps involved in the classification were as follows:

1) Creating train-test split and cross-validation: The original sample was split into 10 equal sized sub-samples. Of the ten sub-samples, one sample was retained as the testing sample and the remaining 9 samples were used for training. This process was repeated 10 times, using each of the 10 samples as the validation set exactly once.

2) Oversampling the minority class: The dataset provided was imbalanced, i.e., there were very few SES and SIS labelled posts, making them the minority classes. Fitting models on imbalanced data results in poor classification results. Hence, it is important to create a balanced training set. In order to do this, the SES and SIS posts were first combined into a broader "Seeking Support" (SS) minority class. Synthetic Minority Oversampling Technique (SMOTE) was then used to over-sample the minority class and under-sample the majority classes, resulting in a balanced dataset. SMOTE was performed only on the training set at each step of the cross-validation to maintain a "clean", unaltered set for testing.

3) Training the classifiers: Each of the classification models mentioned previously were fit on the training sub-set in each fold of the cross-validation.

4) Predicting class labels: The trained classifier was used to predict the class label on the test subset at each stage of cross-validation - '1' if the post belongs to the social support category on which the classifier was trained and '0' if it does not.

5) Evaluating the classifier: ROC-AUC was used as the metric to decide the best performing classifier. The ROC-AUC score was calculated by comparing the target and predicted labels for the test set. The average score across the 10 folds of cross-validation was used to evaluate the classifier performance.

The results (as seen in section 3.6 of this thesis) showed that Random Forest was the best performing classifier. Hence, the Random Forest classifier, with SMOTE was used to classify the rest of the user posts in the dataset, for each of the categories of social support.

3.7 RESULTS AND DISCUSSION

This section highlights the outcomes of the natural language processing and text classification tasks of the project and specifically discusses some of the more noteworthy results and findings. The results of each of the classification models built using Word2Vec for feature embedding were compared against the results from a previous implementation that made use of Bag of Words for feature extraction [2].

10-fold cross validation and oversampling of the minority classes (SES and SIS) were done in both the Word2Vec as well as Bag of Words implementations.

Table 3.2 contains the comparison of the results for the Naïve Bayes classifier:

Label	Word2Vec		Bag of Words	
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.62	0.75	0.7	0.84
PES	0.78	0.68	0.71	0.82
PIS	0.82	0.89	0.75	0.82
SS	0.64	0.7	0.89(SES) 0.85(SIS)	0.75(SES) 0.89(SIS)

Table 3.2: Results of Naive Bayes classification

Table 3.3 consists of the comparison of the results of the Logistic regression model for classification:

Label	Word2Vec		Bag of Words	
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.77	0.88	0.78	0.81
PES	0.85	0.88	0.83	0.79
PIS	0.85	0.92	0.81	0.83
SS	0.82	0.82	0.9(SES) 0.88(SIS)	0.87(SES) 0.8(SIS)

Table 3.3: Results of Logistic Regression for classification

Table 3.4 consists of the comparison of the results of the Support Vector Machine(SVM) model for classification:

Label	Word2Vec		Bag of Words	
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.74	0.88	0.78	0.77
PES	0.85	0.89	0.84	0.68
PIS	0.84	0.91	0.82	0.78
SS	0.83	0.82	0.97(SES) 0.94(SIS)	0.66(SES) 0.75(SIS)

Table 3.4: Results of the SVM model for classification

Table 3.5 consists of the comparison of the results of the Random Forest classifier:

Label	Word2Vec		Bag of Words	
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.84	0.91	0.77	0.85
PES	0.87	0.91	0.83	0.82
PIS	0.84	0.92	0.77	0.84
SS	0.86	0.82	0.97(SES) 0.93(SIS)	0.85(SES) 0.86(SIS)

Table 3.5: Results of the Random Forest classifier

Table 3.6 lists out the results of the Decision Tree classification model:

Label	Word2Vec		Bag of Words	
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.75	0.73	0.77	0.75
PES	0.76	0.71	0.81	0.69
PIS	0.78	0.77	0.78	0.72
SS	0.77	0.6	0.96(SES) 0.94(SIS)	0.67(SES) 0.77(SIS)

Table 3.6: Results of the Decision Tree classifier

Table 3.7 provides a comparative study of the results obtained using Adaboost classification:

Label	Word2Vec	Bag of Words		
	Test Acc.	ROC AUC	Test Acc.	ROC AUC
COM	0.77	0.84	0.8	0.85
PES	0.76	0.72	0.82	0.82
PIS	0.79	0.84	0.8	0.86
SS	0.81	0.73	0.96(SES) 0.91(SIS)	0.67(SES) 0.87(SIS)

Table 3.7: Results of Adaboost classification

The goal of the text classification task in this project was two-fold. Firstly, to investigate the performance of Word2Vec for feature embedding, in comparison to the previously implemented Bag of Words model, since Word2Vec is a more effective tool for feature extraction in problems where semantics are of importance. Secondly, the aim was to find the best performing classifier to categorize the user posts in the dataset, into the different

social support categories. Table 3.5 reveals that Random Forest is the best performing classification model and that the model, when based on Word2Vec does better than a BOW-based model. Hence, Word2Vec was used for feature embedding and Random Forest was used as the classifier of choice to classify the user posts into their respective social support categories.

The classification also helped identify that Providing Informational Support (PIS) is the most popular type of social support activity in the dataset, as seen in Table 3.8.

Type of social support	Number of users in the dataset
COM	689010
PES	816559
PIS	1693524
SS	119304

Table 3.8: Number of users for each social support type

CHAPTER 4

CHURN PREDICTION

This chapter describes the various tasks that were performed in an effort to analyze user engagement in the OHC and predict potential churn of users from the community. The aim of this experiment was two-fold. Firstly, it aimed at analyzing the relationship between the different types of social support activity and user churn. Additionally, it analyzed the effect of studying previous user activity to predict future churn. Some of the significant results obtained from the above analyses have also been discussed towards the end of this chapter.

4.1 DEFINITION OF TRAJECTORY

Yuan et al. define trajectory as follows [25]:

"A trajectory (TR) is a chronological sequence consisted of multi-dimensional locations, which is denoted by $TR_i = \{P_1, P_2, \ldots, P_m (1 \leq i \leq n)$. $P_j (1 \leq j \leq m)$ sampling point in TR_i is represented as <Location j, T_j>, which means that the position of the moving object is Location j at time T_j. Location j is a multi-dimensional location point."

The trajectory, for the purpose of this thesis was defined as a chronological sequence of user activity, which records the change in the type and frequency of social support activity of users in the community, over a period of six months. The trajectory TR, denoted by <Location j, T_j> for each user, consisted of six sampling points corresponding to six months of user activity. Six months is a significant amount of time to understand patterns in user engagement. Furthermore, it provides a reasonably large user pool for analysis. Hence, trajectories representing the evolution of user activity in their first six months, were used to predict churn in the following months.

4.2 BUILDING THE USER POOL

The user pool used for analysis consisted of users who had been active in the community for six months or more and whose last instance of activity was within the last six months of the entire dataset. This user pool was picked keeping in mind the goal of predicting

user churn from previous user activity. Four different churn prediction models were built, which used the first six months of a user's activity to predict user churn in the 7^{th}, 7-8^{th}, 7-9^{th} and 7-10^{th} months respectively. Hence, users with activity only in the last six months of the dataset had to be excluded from the analysis since there is no information beyond the six-month period for such users.

Excluding users who had less than six months of activity and those who had activity in the last six months of the dataset narrowed down the user pool to 12,900 users from the total of 50,000 users in the dataset, which is still a sufficiently large pool of users for analysis.

4.3 THE TRAJECTORY- BASED METHODS

This section of the thesis discusses the three trajectory-based methods that were proposed to predict user churn in the four churn prediction models as described in the previous section. The method hypothesizes that analysis of the trajectory of a user over an extended period of time as well as the type of social activity that the user is engaged in, are important estimators of user churn.

The three trajectory-based methods are described as follows:

1) 5-Element Transitional Trajectory (5-ETT) [26]: In this method, the user activity in each of the six months in the trajectory was represented as a 5-element vector, indicating the number of posts made by the user, corresponding to each of the social support types, in each month. This trajectory-based model differentiates between the different social support types and was based on the concept of state transitions, where the "state" was the 5-element vector representation of user activity and the transition records the change in state from one month to the next. For example, a vector representation of [2, 3, 1, 4, 0] in the first month for a user would indicate that the user made 2 Companionship (COM) posts, 3 Providing Emotional Support (PES) posts, 1 Providing Informational Support (PIS) post, 4 Seeking Support (SS) posts and 0 posts not belonging to any of the types (Oth), in his first month of activity.

2) 1-Element Transitional Trajectory(1-ETT) [26]: This method did not differentiate between the social support types and user activity in each month was represented as a 1-element vector which indicated the sum total of posts made by the user in each month. For example, if the 5-element user activity vector in the first month is given by [2, 3, 4, 1, 0], the 1-ETT equivalent of this state will be given by [2+3+4+1+0] = [10]. This method is similar to the 5-ETT method in that it keeps a track of state transitions from month to month.

3) Consolidated Trajectory: This method combines the 5-element user activity vectors across the six months into a single 30-element long vector. The trajectories in this method treat the different social support types separately but do not have a transitional component.

4.4 CHURN PREDICTION

The churn prediction models for the transitional trajectory models (5-ETT and 1-ETT) were built using a Bayesian approach while the Consolidated Trajectory model used a Logistic Regression classifier. The steps for sample extraction and model setup are as follows:

1) Building the churn, non-churn samples and labels: The first step in setting up the prediction models to build features for the models from the existing user pool of 12,900 users. The trajectories of the first six months of activity for each of the 12,900 users were extracted as input samples. Once the samples were built, target labels were assigned to each of the samples, separately for the four churn prediction models i.e labels '1' or '0' were allotted depending on 'no churn' or 'churn' between the 6^{th} and $(6+k)^{th}$ month, where $k \epsilon [1,4]$

2) Binning the samples: The next step was to bin the user activity in the churn and non-churn samples obtained above into "Low", "Medium" and "High" to reduce the sparsity across the trajectories. The thresholds for the three bins were set separately for each of the social support categories and each month in the user trajectory. For example, the binning thresholds for COM posts across all users in the 1^{st} month

would be different from those for COM posts in the 5th month. Also, binning thresholds for COM and PES posts would be different. This is because the frequency of posts of a certain kind of social support could be higher or lower than the rest – for example, SS posts were a minority class in our dataset. Hence, the model could not have uniform binning thresholds across the different social support categories. Analysis of histograms of user activity across months in the dataset showed a monotonically decreasing trend. Therefore, the binning thresholds for each month also had to be set independently of the other months.

An example of the binning is as follows:

Given an initial user activity vector of a given month as

[COM, PES, PIS, SS, Oth] = [10, 0, 3, 0, 3]

After binning

[COM, PES, PIS, SS, Oth] = [2, 0, 0, 0, 1]

where the labels '0'indicates the 'Low' bin, '1' indicates the 'Medium' bin and '2' stands for the 'High' bin. The above example demonstrates that a posting frequency of 3 for PIS posts was considered 'Low' while that for Oth was considered 'High'. The thresholds of the bins were set such that such that a post frequency of 0 was treated as 'Low' across all social support types and across all months in the trajectory. The median of posting frequency for each social support type, for each of the six months was used to set the 'Medium' and 'High' binning thresholds- posting frequency less than the median was binned 'Medium' and that greater than the median was binned 'High'.

3) Random Sampling and 10-fold Cross Validation: The original samples were split into 10 equal sized sub-samples. Of the ten subsamples, one sample was retained as the testing sample and the remaining 9 samples were used for training. The dataset used for this task was also imbalanced, given the low rate of user churn from month to

month, resulting in fewer churn samples for prediction. This imbalance was overcome by under-sampling the non-churn samples in the training set.

Once the samples were extracted and pre-processed, they were used as inputs to a Bayesian model and a Logistic Regression model, depending on the type of trajectory model used.

The Bayesian implementation for the 5-ETT and 1-ETT models was based on the traditional Bayes' Theorem and is described as below:

$$P(d=1|Tr) = \frac{P(Tr|d=1).\ P(d=1)}{P(Tr|d=1).\ P(d=1) + P(Tr|d=0).\ P(d=0)} \quad \textbf{(4.1)}$$

$$P(d=0|Tr) = \frac{P(Tr|d=0).\ P(d=0)}{P(Tr|d=1).\ P(d=1) + P(Tr|d=0).\ P(d=0)} \quad \textbf{(4.2)}$$

where 'd' indicates the destination of the user, which in this case was either churn(d=0) or no churn (d=1) from the OHC. 'Tr' denotes the six-month trajectory of the user. P(Tr|d=1) and P(Tr|d=0) are the likelihoods of a trajectory belonging to a "non-churn" or "churn" user respectively. The priors, P(d=1) and P(d=0), are the marginal probabilities of "non-churn" and "churn: users respectively. P(d=1|Tr) and P(d=0|Tr) are the posterior probabilities of no churn and churn given a user trajectory.

Since, this model is based on transitional trajectories, transition matrices built separately for non-churn and churn users were used to obtain the likelihoods P(Tr|d=1) and P(Tr|d=0). These matrices help record the state transitions in the trajectory. A state here indicates either the 5-element or 1-element user activity vector, depending on the transitional trajectory model. Each of the states in the trajectory was assigned a unique state number, for easy lookup in the transition matrix. The rows in the transition matrix represented the state of the user in Month i, (S_i) and the columns in the matrix represented the state of the user in Month (i+1), (S_{i+1}). The entry corresponding to (S_i, S_{i+1}) in the matrix represented the probability of transition from State S_i to State S_{i+1} across all users in the training set. Each trajectory consists of 5 state transitions; hence the total likelihood of a trajectory is the product of the probabilities of all the state transitions in it.

A mathematical representation of the total likelihood of the trajectory is given as:

$$P(Tr|d) = \prod_{i=1}^{5}(S_i, S_{i+1}) \quad \textbf{(4.3)}$$

For the 5-ETT method, the size of the transition matrix was 3^5 x 3^5, given that each of the five types of social support contained in the vector could take 3 possible values 0, 1 or 2, depending on the bins assigned to them. The 1-ETT method had only a single number representing state, thus resulting in a transition matrix of size 3 x 3 for this method.

The posterior probabilities, P(d=1|Tr) and P(d=0|Tr) were used to predict churn in the four churn prediction models. P(d=0|Tr) > P(d=1|Tr) indicated churn, while a higher value of P(d=1|Tr) indicated non-churn.

The 30-element long consolidated trajectories were given as input to a Logistic Regression Classifier. The output of the classification was a probabilistic estimate of each trajectory belonging to the positive i.e. non-churn class. This estimate was used to predict churn- a value greater than 0.5 meant non-churn else churn.

Several metrics like accuracy, ROC-AUC, Precision and Recall were used to evaluate each of the trajectory-based methods across the four churn prediction models.

4.5 BENCHMARKS

While evaluation metrics help measure the individual performance of machine learning models, it is important to compare the performance of the model against other similar models to gain a measure of confidence to assess which of the methods will best achieve the objectives of the modelling exercise. The three trajectory-based methods were compared against two non-trajectory-based benchmarks, to test the effectiveness of trajectory-based models for predicting churn.

1) Benchmark 1: This benchmark proposed that churn behavior in the $(6+k)^{th}$ month can be predicted solely based on the user activity vector in the 6^{th} month. Assigning target labels, binning and random sampling was done in a manner similar to the trajectory-

based methods. Logistic Regression was used to get a probabilistic estimate to eventually predict churn and the same evaluation metrics were calculated.

2) Benchmark 2: This benchmark was similar to Benchmark 1, differing only in that it used the sum total of user activity in the 6^{th} month to predict churn behavior in the $(6+k)^{th}$ month i.e. it did not differentiate between the social support types. Similar to Benchmark 1, this benchmark also did not take user trajectory into consideration and used Logistic Regression.

4.6 RESULTS AND DISCUSSION

This section provides a comparative study of the results of the proposed trajectory-based methods and the two non-trajectory-based benchmarks for the four churn prediction models. Tables 4.1- 4.4 enlist the different performance metrics across all methods for the different churn prediction model.

1) Model 1: Churn Prediction in Month 7, d=0: 736, d=1: 12164

Metric	5- ETT	1- ETT	Consolidated	Benchmark 1	Benchmark 2
Accuracy	0.66	0.65	0.79	0.77	0.72
Precision d=1	0.98	0.97	0.99	0.99	0.94
Recall d=1	0.65	0.65	0.78	0.76	0.76
Recall d=0	0.79	0.69	0.83	0.82	0.14
Precision d=0	0.12	0.10	0.19	0.17	0.03
AUC	0.73	0.68	0.86	0.8	0.59

Table 4.1: Results of churn prediction in the 7^{th} month

2) **Model 2: Churn prediction between 7th and 8th month, d=0: 1453, d=1: 11447**

Metric	5- ETT	1-ETT	Consolidated	Benchmark 1	Benchmark 2
Accuracy	0.55	0.42	0.75	0.75	0.69
Precision d=1	0.93	0.93	0.93	0.93	0.89
Recall d=1	0.53	0.39	0.78	0.77	0.74
Recall d=0	0.70	0.77	0.56	0.55	0.27
Precision d=0	0.16	0.14	0.24	0.23	0.12
AUC	0.65	0.64	0.7	0.64	0.51

Table 4.2: Results of churn prediction between the 7th- 8th month

3) **Model 3: Churn prediction between 7th and 9th month, d=0: 2051, d=1: 10849**

Metric	5- ETT	1-ETT	Consolidated	Benchmark 1	Benchmark 2
Accuracy	0.43	0.41	0.69	0.73	0.43
Precision d=1	0.88	0.89	0.89	0.88	0.83
Recall d=1	0.37	0.35	0.72	0.78	0.4
Recall d=0	0.75	0.78	0.53	0.45	0.56
Precision d=0	0.19	0.18	0.26	0.28	0.15
AUC	0.61	0.61	0.67	0.6	0.49

Table 4.3: Results of churn prediction between the 7th- 9th month

4) Model 4: Churn prediction between 7th and 10th month, d=0: 2609, d=1: 10291

Metric	5- ETT	1-ETT	Consolidated	Benchmark 1	Benchmark 2
Accuracy	0.44	0.44	0.63	0.7	0.45
Precision d=1	0.85	0.86	0.85	0.84	0.79
Recall d=1	0.36	0.35	0.66	0.78	0.42
Recall d=0	0.75	0.76	0.55	0.4	0.56
Precision d=0	0.23	0.23	0.29	0.32	0.2
AUC	0.6	0.59	0.63	0.72	0.51

Table 4.4: Results of churn prediction between the 7th- 10th month

The purpose of building four churn prediction models was to be able to analyze the extent of effectiveness of the six-month trajectory to predict churn in the $(6+k)^{th}$ month, $k \epsilon [1,4]$. Tables 4.1-4.4 show that there is a generally decreasing trend across parameters for lower values of k and the parameters start becoming inconsistent for higher values of k. This indicates that the six-month trajectory is effective in predicting churn only in months immediately following the trajectory. The trajectory becomes noise as k increases i.e. as the period of churn becomes larger to include months that are far from the 6^{th} month.

The discussion of the results will hence be limited only to Models 1 and 2 i.e. churn prediction in the 7^{th} month and the 7-8^{th} month. Tables 4.1 and 4.2 show that the 5-ETT and Consolidated Trajectory models, which differentiate between the social support categories perform better than the 1-ETT trajectory-based model which does not distinguish between the different types of social support. A similar observation can be drawn by comparing the results of Benchmark 1 and 2. Thus, indicating that the different social support types individually affect churn.

The Bayesian implementation for the transitional trajectory-based models faces the problem of "unseen" trajectories in the testing set because of the sparse nature of the transition matrices. A trajectory is said to be "unseen" when the probabilities of one or more of the state transitions in the trajectory are not present in the transition matrices. The 5-ETT method could have 3^5 unique states, resulting in many unique trajectories. Thus, resulting in an increased possibility of encountering unseen trajectories in the testing set. The 1-ETT model would have a lower number of unseen trajectories when compared to the 5-ETT model. However, the Consolidated Trajectory model uses Logistic Regression and hence, solves the problem of having unseen trajectories. Hence, the Consolidated trajectory model was chosen for comparison against the non-trajectory-based benchmarks.

Comparing the results of the Consolidated Trajectory model with Benchmark 1 and 2 shows that the trajectory-based model does better than the non-trajectory-based models and are hence worth exploring in the context of churn prediction.

Recall and precision for the d=0 or churn class are important metrics to evaluate the performance of the methods since the goal churn prediction or prediction on the d=0 class. A close look at the parameters would show a low precision for the churn class across all models. This is because the test set was unaltered and thus imbalanced. Undersampling was done only on the training set at each fold of cross-validation.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

The results of classification and trajectory mining provided important information regarding user participation in the community. Classification of the user posts showed that the most popular social support activity in the OHC was providing informational support. This is expected because the community largely consists of survivors of a disease and exchanging information about the disease is why users join the OHC to begin with. Correlation studies also showed that providing information was positively correlated with user retention in the community while seeking information was negatively correlated i.e. users who are in the community for seeking information may not stay on in the OHC [2].

Comparison of the results across the three trajectory-based models and the benchmarks also further reinforces the proposition that the various social support activities affect churn differently. The key conclusion from the results of the churn prediction was that studying the user activity over a period of time using trajectories is effective in the prediction of future activity of the user in the OHC. This is because users may tend to post frequently when in need of support or while providing information about a situation they closely relate to. There might also be periods of inactivity when users do not post in the community at all. Hence, prediction of future user engagement based on an isolated month of activity is not very accurate. A trajectory helps keep a record of spurts of high activity, moderate activity, and even inactivity, thus providing a more comprehensive representation of user activity.

One of the biggest challenges throughout the project was the imbalanced nature of the datasets, having very few instances for some of the social support types in the text classification task and very churn samples in the churn prediction task. Rebalancing the instances in the datasets will potentially make the text classification and methods involving Bayesian implementations more robust. Another limitation of the study was

that it was based on data obtained only from one breast cancer OHC. The social support patterns and user engagement in OHCs for other diseases and conditions may vary. That said, the methods proposed in this thesis to classify social support and analyze user engagement can be applied to other OHCs.

5.2 FUTURE WORK

There are several interesting avenues for further research. An immediate extension of the current study would be to explore an ensemble approach using all three trajectory-based methods. Since it has been concluded that it is important to distinguish between the different social support types to predict churn, it would be exciting to investigate the effect of each or different combinations of the social support categories, on churn. Another avenue worth considering would be to go beyond the problem of churn prediction and use trajectory-based methods to detect short-term and long-term engagement users in the community.

Furthermore, Trajectory Clustering techniques could be used to find similarities or patterns in participation across users in the OHC. Detecting the health status of the user could also be a valuable path for research, to understand why users leave the community and work on efforts to retain users. Yet another interesting effort would be to use a balanced dataset in order to build an integrated prediction model that is applicable across all OHCs.

The results of this study can serve as useful pointers for OHC administrators to sustain user participation and facilitate user retention through interventions like recommendations, alerts etc. A practical implication of the churn prediction model could be to identify users who are likely to leave the community, thus enabling more targeted intervention [2].

REFERENCES

1. Breast Cancer Information and Awareness. (n.d.). Retrieved from http://www.breastcancer.org/

2. Wang, X., Zhao, K., & Street, N. (2014). Social Support and User Engagement in Online Health Communities. *Smart Health Lecture Notes in Computer Science*, pp. 97–110., doi:10.1007/978-3-319-08416-9_10.

3. Sharma, N. (2015). Patient centric approach for clinical trials: Current trend and new opportunities. *Perspectives in Clinical Research,6*(3), 134. doi:10.4103/2229-3485.159936

4. Fox, S. (2011). The Social Life of Health Information. *Pew Research Center: Internet, Science & Tech*, www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

5. Swan, M. (2012). Crowdsourced Health Research Studies: An Important Emerging Complement to Clinical Trials in the Public Health Research Ecosystem. *Journal of Medical Internet Research,14*(2). doi:10.2196/jmir.1988

6. Leiter, A., Sablinski, T., Diefenbach, M., Foster, M., Greenberg, A., Holland, J., & Galsky, M. D. (2014). Use of Crowdsourcing for Cancer Clinical Trial Development. *JNCI: Journal of the National Cancer Institute,106*(10). doi:10.1093/jnci/dju258

7. Burrows, R., Nettleton, S., Pleace, N., Loader, B., & Muncer, S. (2000). Virtual Community Care? Social Policy and the Emergence of Computer Mediated Social Support. *Information, Communication & Society*, vol. 3, no. 1, pp. 95–121., doi:10.1080/136911800359446

8. Qiu, B., Zhao, K., Mitra, P., Wu, D., Caragea, C., Yen, J., Greta, G. E., & Portier, K. (2011). Get Online Support, Feel Better -- Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. *2011 IEEE Third Intl Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Intl Conference on Social Computing*, doi:10.1109/passat/socialcom.2011.127

9. Dunkel-Schetter, C. (1984). Social Support and Cancer: Findings Based on Patient Interviews and Their Implications. *Journal of Social Issues*, vol. 40, no. 4, pp. 77–98., doi:10.1111/j.1540-4560.1984.tb01108.x

10. Rodgers, S., & Chen, Q. (2005). Internet Community Group Participation: Psychosocial Benefits for Women with Breast Cancer. *Journal of Computer-Mediated Communication,10*(4), 00-00. doi:10.1111/j.1083-6101.2005.tb00268.x

11. Shumaker, S. A., & Brownell, A. (1984). Toward a Theory of Social Support: Closing Conceptual Gaps. *Journal of Social Issues,40*(4), 11-36. doi:10.1111/j.1540-4560.1984.tb01105.x

12. Kraut, R., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., Konstan, J., Ren, Y., & Riedl, J. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, MA: The MIT Press

13. Young, C. (2013). Community Management That Works: How to Build and Sustain a Thriving Online Health Community. *Journal of Medical Internet Research,15*(6). doi:10.2196/jmir.2501

14. Iriberri, A., & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Computing Surveys,41*(2), 1-29. doi:10.1145/1459352.1459356

15. House, J. S. (1983). *Work stress and social support*. Reading, MA: Addison-Wesley.

16. Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl, A., & Birch, G. E. (2008). Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. *2008 Seventh International Conference on Machine Learning and Applications*. doi:10.1109/icmla.2008.34

17. Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research,43*(2), 276-286. doi:10.1509/jmkr.43.2.276

18. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data

mining approach. *European Journal of Operational Research,218*(1), 211-229. doi:10.1016/j.ejor.2011.09.031

19. Datta, P., Masand, B., Mani, D.R. et al. (2000). Automated Cellular Modelling and Prediction on a Large Scale. *Artificial Intelligence Review,14: 485*. doi:10.1023/A:1006643109702

20. Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks,11*(3), 690-696. doi:10.1109/72.846740

21. Natural Language Toolkit. (n.d.). Retrieved from https://www.nltk.org/

22. Gensim: Topic modelling for humans. (n.d.). Retrieved from https://radimrehurek.com/gensim/models/word2vec.html

23. Compute Engine – IaaS. Google Cloud. (n.d.). Retrieved from https://cloud.google.com/compute/

24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Jeffrey, D. (2013). Distribution Representation of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems 26*, 3111-3119

25. Yuan, G., Sun, P., Zhao, J., Li, D., & Wang, C. (2016). A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review,47*(1), 123-144. doi:10.1007/s10462-016-9477-7

26. Xue, A. Y., Zhang, R., Zheng, Y., Xie, X., Huang, J., & Xu, Z. (2013). Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. doi:10.1109/icde.2013.6544830