Predicting User Posting Activities in Online Health Communities with Deep Learning

Xiangyu Wang, Kang Zhao[1], Xun Zhou, and Nick Street
The University of Iowa
{xiangyu-wang-1, kang-zhao, xun-zhou, nick-street}@uiowa.edu

ABSTRACT

Online Health Communities (OHCs) represent a great source of social support for patients and their caregivers. Better predictions of user activities in OHCs can help to improve user engagement and retention, which are important to manage and sustain a successful OHC. This paper proposes a general framework to predict OHC users' posting activities. Deep learning methods are adopted to learn from users' temporal trajectories in both the volumes and content of posts published over time. Experiments based on data from a popular OHC for cancer survivors demonstrate that the proposed approach can improve the performance of user activity predictions. In addition, several topics of users' posts are found to have strong impact on predicting user's activities in the OHC.

KEYWORDS

Predictive model, Trajectory mining, Text analytics, User churn

Citation

Xiangyu Wang, Kang Zhao, Xun Zhou, and Nick Street. 2020. Predicting User Posting Activities in Online Health Communities with Deep Learning. *ACM Trans. Manage. Inf. Syst.* (11:3), Article 12 (July 2020), 15 pages. https://doi.org/10.1145/3383780

1 Introduction

Nowadays more and more people use information technologies to take control of their own or family members' health. In the United States, 80% of adult Internet users use the Internet for health-related purposes. Among them, 34% read health-related anecdotes or comments from others [9]. Online health communities (OHCs) represent a popular way for those with similar health concerns or problems to interact and network with each other online. The use of OHCs has made communication multi-directional, information more easily accessible and problem sharing/solving simpler and faster [9,15,30]. The benefits of getting involved in OHCs have been well documented in the literature, such as empowerment [4], reduced stress [27], and improved offline life [18]. More importantly, these benefits are not limited only to those who receive support from OHC -- actively helping others is beneficial to support providers as well [8].

Nevertheless, as in many other online communities, there are great variations in the levels of activities among OHC users -- many users simply lurk without posting anything, while those who have posted can often be inactive in posting or even leave the community. While lurking can help individual users too [46], the key for the success of an OHC lies in providing assistance to its members and active participation of as many members as possible -- these are not possible if users do not post any content. Therefore, better understanding and accurate predictions of users' posting activities from OHCs can help to build and sustain a successful OHC, which will eventually benefit OHC users, through improved community design, management, and user retention [34]. For instance, with an accurate activity-prediction model, an OHC can identify users who are likely to be inactive for an extended period of time and send emails to remind them of ongoing discussions or new users in need of help in the OHC.

In this paper, we predict OHC users' future posting activities by learning from their temporal trajectories in volumes and content of their posts. The contributions of this paper are threefold: First, from a design science perspective, we set up a more general and flexible framework to predict OHC users' posting activities. This framework can accommodate activity predictions for future time periods with various lengths (including the prediction of churns as defined by community managers) and predictions for different types of users, including new users and continued users; Second, from a methodological perspective, we incorporate user-generated content into the prediction of user posting activities and show that our proposed approach -- parallel LSTM with Topics -- provides better predictions than using trajectories of posting volume only; Third, from a managerial perspective, with the help of the predictive model, we reveal several topics that OHC users tend to discuss before they become inactive or keep active, and thus offer insights into the design of user retention interventions.

[1] Corresponding author: S224 PBB, Iowa City, IA 52242.

The remainder of the paper is organized as follows: we review related work in Section 2, followed by describing our methods and showing results in Section 3. Section 4 reveals topics related to users' post activities. The paper concludes with discussions of limitations and future research directions.

2 Related Work

Studies on patient support groups have a long history, but earlier studies relied mainly on data collected through questionnaires or interviews. Compared to traditional offline support groups, OHCs not only enable but also record distributed online interactions among individuals. This makes large amounts of data available for computational analysis, making it possible to study individuals' behaviors online in a real-time fashion. During the past a few years, researchers have used computational methods to mine OHC datasets to study adverse drug reactions [41], social network dynamics [46], social support patterns [42], user roles [34,45], language use [37], sentiment towards treatment [24], content recommendations [40], offline health outcomes [33], etc. A recent review of studies in this area can be found in [38].

Users' engagement in an OHC is closely related to the content of their posts. Many studies have examined such user-generated content to better understand users' online behaviors, mainly with semantic features and syntactic features. Commonly used semantic features include N-gram (e.g. bigram, trigram and bag-of-words) [5,17,33] and dictionaries (e.g. Linguistic Inquiry and Word Count (LIWC) and Named Entity Recognition (NER)) [6,27,36–38,43,45]. Commonly used syntactic features include parts-of-speech (POS) [36] and dependency trees (e.g., parsing trees). Besides these linguistic features, topic modeling techniques such as Latent Dirichlet Allocation [2] have been used to detect latent topic features from UGC, including identifying types of social support and mining meanings from post contents in the OHCs [6,25,34,36]. Recent studies on health-related UGC used word embeddings (e.g., Word2Vec) [21] to capture both semantic and syntactic characteristics [14,43,47]. Word2Vec and LDA have also been combined to identify debates in an online breast cancer community [44].

User churn is a common issue for many online communities and social media, and thus has been address with computational approaches [29]. Researchers have attempted to incorporate social network structures among users [16], different types of user activities (e.g., asking vs answering a question) [26], users clusters/types [7,22,28,39] and the diffusion of churn behaviors between users [35]. Specifically, for OHCs, Wang et. al developed a model to predict OHC users' churn behaviors, which were defined as a long time period of inactivity [35]. The model designed a few features to capture how users' number of posts change overtime and added each post's social support types.

Despite extensive research in churn prediction, there are still a few limitations: First, the definition of "churn" is usually ad-hoc. While user churn in certain areas can be clearly defined, such as cancelling cable TV services, it is more challenging to define churn in online communities or services--they are mostly free to join, and there is no formal procedure for users to declare their departure. Also, even after a long hiatus, some users may come back. In the literature, the length of inactivity to qualify as "churn" for online communities varies from studies to studies and can range from 1 month to 1 year. Second, most studies focused on new users only. They observed new users' first a few weeks/months, or the first a few online activities (e.g., posts), and then made predictions on whether they will churn. While this is useful, many online communities, including OHCs, may want to expand such predictions to all users, especially long-term users whose experience can be more valuable to the community. Third, many of the previous studies need significant amount of manual feature engineering or annotations. On one hand, many focused on using snapshots of users' online activity profiles for churn predictions. Among those that did consider trajectories of behaviors, features need to be manually designed to capture temporal changes in users' behaviors as in [26,35]. On the other hand, few have leveraged user-generated content, a valuable source of data that is unique to online communities and social media. While some have considered simple characteristic (e.g., length of posts), the rich data embedded in content is rarely used for churn prediction, with the exception of [35]. However, their text classifiers for social support types need to be trained by human annotated datasets, and thus have limited applicability in OHCs for other diseases or online communities in other domains.

Therefore, in this research, we propose to predict if an OHC user will post anything during the next M months. Thus, by adjusting the value of M, an OHC can better decide the appropriate length of user inactivity before it should intervene instead of picking an arbitrary definition of "churn". Such a prediction is based on observation of users' activities during a time period of K months prior to the prediction, and thus can be done for each user whose has a history of at least K months of online activities. We propose to use LSTM to learn from the trajectories of users' posting volume and content. Post volumes can be directly fed into LSTM without feature engineering, while post content is captured by topic modeling, an unsupervised approach that does not require human annotations.

3 Methods and Results

The dataset used in this paper consists of user-generated posts from Breastcancer.org, a popular peer-to-peer online health community for breast cancer survivors and their caregivers. It comprises all public posts from this community between October 2002 and August 2013. There are more than 2.8 million user posts from 107,549 threads, contributed by around 50,000 users.

In our dataset, we found that a long period of hiatus does not necessarily mean a user will leave the OHC: 20.9% of the users who were inactive in posting for 3 months in the OHC came back later to publish a post. Similarly, the conditional probability of a user's return to posting given 6 months of inactivity is still 13%. This echoes our previous discussions on the difficulty of defining user "churn" for OHCs, or online communities in general. Therefore, our prediction will observe users' posting histories during the past *K* months and decide if they will post anything during the next *M* months, with different *K* and *M* values. Our approach uses post volumes first, and then adds post content as features to illustrate their contributions to the prediction.

3.1 The trajectory-based method to predict user activities

3.1.1 Capturing User Activity Trajectory.

As we mentioned earlier, users will come back again even after a long period of hiatus. Thus, we first try to predict users' posting in the next month (*M=1*) by observing user's behaviors during the past *K=6* and *K=12* months respectively. We define that a user is active in posting (labeled as Class 1) during month t if she publishes one or more posts during that month. Otherwise, the user would be labeled as inactive during the month (class 0).

A user's posting activities during past *K* months constitute a trajectory. It can be represented with a time series with *K* elements, each element representing the number of posts the user published during each of the *K* months. For example, when trying to predict a user's posting activity during month *t* with an observation period of *K=6* months, a post volume trajectory <4, 6, 0, 8, 2, 0> indicates that the user published 4 posts during the *(t-6)*th month, 6 posts during the *(t-5)*th month, …, and no post during the *(t-1)*th month.

In our approach, a user can have multiple predictions based on when the prediction is made. For example, we can predict if a user will be active during her 13th month based on her activities during her 1st-12th months. We can also predict if the user will post during her 14th month based on her posting activities during her 2nd-13th months. Therefore, our predictive model is based on user- observation window instances. For example, a user whose first post was *N* months away from the end date of our dataset would have *N-(K+M)+1* such user-observation window instances, because we need an observation period of *K* months for each user and predict the next *M* months. This also means we need to remove users whose first post was published within the last *K+M* months of our dataset. For example, with *K=12* and *M=1*, this yields a pool of 41,788 users out of 50,414 users in the OHC with 1,504,938 instances.

Since we observe each user's posting activities from her first post till the end date in our dataset using a sliding time window, another problem in our dataset is the existence of a large number of instances with all zeroes. This is caused by users who have long hiatus and then return to the OHC and post again. For example, a user who was inactive in posting for 12 months and then return to post would have 12 consecutive zeroes in her posting volume trajectory. With *K=6* and *M=1*, these 12 zeroes in the trajectory will lead to 7 instances with all zeroes. Meanwhile, in our user pool, most of these all-zero instances belong to the negative class (i.e., inactive): 99.4% when *K=6* and *M=1*, and 99.7% when *K=12* and *M=1*. Therefore, to simplify the model, we exclude those all-zero instances from our model as simply predicting them to be negative without learning can already get very good results. Including them in the learning process would lead to a more unbalanced dataset.

3.1.2 Incorporating the Long Short-Term Memory (LSTM) Neural Network model.

The Long Short-Term Memory Neural Network model [12] is generally considered to have strong capability to learn from temporal or sequence data. LSTM has two states: one is cell state $c^{(t)}$, another is a hidden state $h^{(t)}$. The Memory cell, modulated by three gates – input, output and forget gates – is the major functional component of the LSTM. These gates determine the amount of dynamic information entering/leaving the memory cell. The memory cell has a set of internal states, which store the information obtained over time [32]. These internal states constitute a representation of an input sequence learned over time. Specifically, the input gate controls the degree to which the input information would enter the memory cell to influence its hidden state h_t at time t. The forget gate modulates the previous hidden state h_{t-1} to control its contribution to the current state. The output gate gets the information output from a memory cell which would influence the future states of LSTM cells.

The trajectory-based LSTM unit contains a memory cell that stores long-term user activity trajectory information. The structure of an LSTM cell used in [12], as shown in the Figure 1, at each time step, this memory cell takes a new user monthly activity trajectory record $x^{(t)}$ and maintains a portion of prior user activity information $h^{(t-1)}$. The trajectory-based LSTM unit also contains an input gate $i^{(t)}$, a forget gate $f^{(t)}$ and an output gate $o^{(t)}$. The results from activation functions are $z^{(i)}, z^{(f)}, z^{(o)}$. The three gates inherit prior user activity trajectory information and also receive the current user activity trajectory. To be specific, $z^{(f)}$, the forget gate result, controls cell state from last state $c^{(t-1)}$ on how much will enter and how much will leave. $z^{(i)}$, the information gate result keeps selective memory from the input $x^{(t)}$. $z^{(o)}$, the output result determines which parts are going to be outputs at time t.

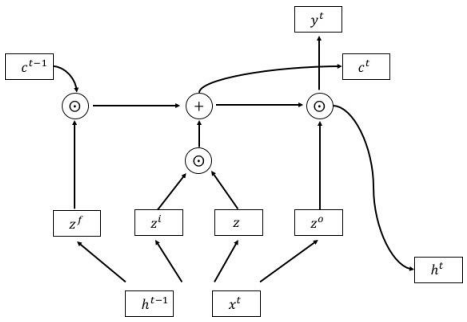

Figure 1: An illustration of the structure of an LSTM model cell.

We first deploy LSTM with one hidden layer and a single unit to learn from user posting trajectories. It was implemented in Keras on top of Tensorflow, along with Scikit-learn on the Argon High Performance Computing system GPU having 56-core 256GB with Titan V accelerators.

3.1.3 Adding Post Content to the Prediction.

People talk about different issues in an OHC, with some discussions not directly related to health issues [35]. We conjecture that what users talk about can also have implications to their subsequent posting behaviors. Thus, we first adopt topic modeling to investigate latent topics from the content of users' posts. After stop words removal and stemming, we apply Latent Dirichlet allocation (LDA) [2]. As a result, each topic is represented by a distribution over keywords and each post has a distribution over these topics.

With LDA, each topic can be represented as a vector that indicates its probability of belonging to each topic. For a user with multiple posts during a month, we can represent the user's discussion topics during that month by averaging topic distributions of her posts during the month. Then over the observation period of K months, each user would have another trajectory on her discussion topics over time. Therefore, our next model adds these new trajectories to the previous LSTM with only post volume trajectories as inputs. For users who have no post during a given month, her topic distribution over that month would be all zeros. As shown in Figure 2, this new model (P-LSTM+Topics) uses two parallel LSTMs, each with one hidden layer and a single unit. The combined outputs of the two parallel LSTMs are fed into one fully connected layer before a prediction is generated.

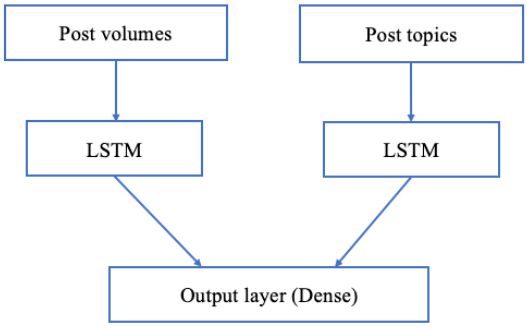

Figure 2: Structure of P-LSTM+Topics, which has two parallel LSTMs, one for post volumes and the other for topics from LDA.

Another method to represent user-generated content by a real-valued vector is word embedding. Word embedding methods [21] have been used to represent words in a low dimension by minimizing the distance between a word and its context words in the vector space.

Then we generate the embedding for a post by averaging embeddings of its words. A user with multiple posts during a month would be represented by the average word embedding of all her posts during the month. For users who have no post during a given month, her post word embedding vector would be an all-zero vector. As shown in Figure 3, the structure of the model based on embeddings (named P-LSTM+Embeddings) is very similar to P-LSTM+Topics (Figure 2).

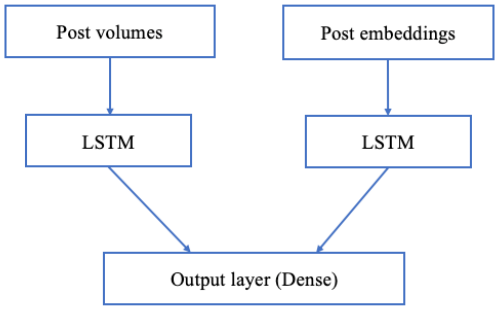

Figure 3: Structure of P-LSTM+Embedding.

3.2 Results

3.2.1 Predictions based only on posting volume trajectories (*K = 6, M =1*).

We first test the performance of different models with *K=6* and *M=1*, i.e., predicting posting activities during the next month after observing users for 6 months. Our dataset for this setting starts with 1,755,666 instances, with a class ratio (inactive vs active) of 17:1. After removing instances with six consecutive zeros, we have 167,004 instances in the inactive class (0) and 88,025 instances in the active class (1), leading to a more balanced class ratio of 2:1.

To illustrate the performance of LSTM, we include four baseline models: Logistic Regression, Decision Tree, Random Forest, and Gated Recurrent Unit (GRU). GRU is another variant of recurrent neural network (RNN) with a simpler architecture and fewer parameters compared with LSTM. To compare the performance of our proposed methods and baselines, we use 10-fold cross validation for each method. The predictive performance is measured by four metrics: Area under the ROC (AUC), accuracy, as well as F1 scores for both the inactive and active classes. Table 1 shows the results, which are the average of 10 times 10-fold cross-validations.

Overall, LSTM outperforms the three baseline models besides GRU in both AUC (all with p-values<0.05 in paired t-tests) and accuracy. LSTM also outperforms GRU, although the difference is not statistically significant. In other words, LSTM can better learn from temporal trajectories of users' post volumes. We also note that for our LSTM with one hidden layer and a single unit, adding more units or layers does not lead to significant improvement in performance.

Table 1. The performances of predictive models (*K=6, M=1*) with best performance in bold.

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
Logistic Regression	0.894	0.827	0.879	0.693
Decision Tree	0.762	0.818	0.867	0.708
Random Forest	0.885	0.837	0.879	0.748
GRU	0.856	0.825	0.878	0.675
LSTM	0.900	0.846	0.887	0.760
P-LSTM+Topics	**0.904**	**0.849**	**0.889**	**0.765**
P-LSTM+Embeddings	0.891	0.837	0.876	0.762

3.2.2 Parallel LSTM with post contents.

For LDA, we try to use 20 and 30 as the numbers of topics for this corpus. To help us better understand the impact of topics in our prediction, we examine and try to interpret each topic generated by LDA. It turns out when there are 30 topics, several topics have overlaps that can be easily observed. Thus, in subsequent analyses, we pick the results based on 20 topics for the ease of interpretation. Table 2 shows the keywords, our interpretations, and an example post for each topic. We will later examine if our results are robust when changing the number of topics. For word embedding, we adopt 300 as the size of embedding vectors.

To compare with predictions based solely on post volume trajectories, we use the same settings of K=6 and M=1. Results with the same evaluations based on 10 times 10-fold cross validations are listed in Table 1.

Results in Table 1 show that P-LSTM+Embeddings does not perform well, as it is dominated by P-LSTM+Topics and LSTM, both with p-values<0.001 in paired t-tests. In other words, although post contents represented by embeddings do not help, adding the trajectories of latent topics from posts can contribute to better predictions (with p-value<0.001 in paired t-test). In addition, topics are generally easier to interpret than embeddings. Therefore, our subsequent analyses focus on P-LSTM+Topics.

3.2.3 The effect of topics on non-LSTM models.

We have shown that adding topic features to LSTM helps to make better predictions (*K=6* and *M=1*). To further illustrate the contributions of topic trajectories, we also add users' trajectories on these 20 topics from LDA to each baseline model. The settings of experiments are the same as previous ones with 10 times 10-fold cross validation.

Similar to LSTM, all models, except Decision Tree, benefit from adding topics. At the same time, P-LSTM+Topics is still the best performer statistically significant (with all p-values<0.001 in paired t-tests). This further highlights the value of adding trajectories of user-generated content in user activity prediction.

3.2.4 Varying the number of topics for LDA models.

To examine whether predictions that incorporate topic trajectories are robust against the number of topics, we try different numbers of topics (K = 10, 30, and 40) and compare results with those based on 20 topics. The experiments are still based on observing user's previous six-month posting activities and predicting the user's next month activity (*K=6* and *M=1*). Results are the average of 10 times 10-fold cross validation.

Results in Table 4 show varying the number of topics in our proposed parallel LSTM does not lead to significant differences in predictive performances measured by AUCs (with all p-values > 0.1 in paired t-tests). In other words, prediction performance is not sensitive to the number of topics in LDA. As a result, we keep our 20 topics as we find that they are easier to interpret than other numbers of topics.

Table 2. Details of the 20 topics from LDA.

Topic	Top keywords	Interpretations	Example post (some content ommitted for brevity)
1	get, eat, like, use, make, food	Healthy recipes for everyday living.	I have been told for health reasons to move away from plain flour and eat things made of whole wheat. I would like to know more about whole-wheat pasta.
2	once, week, chemo, treatmet, scan, herceptin	Pre-treatments discussions.	Is anyone else starting chemo this month? I'll get a thread started, and let me know if you want to be added to this list.
3	get, day, chemo, feel, week, pain	Sharing personal experience and emotion related to treatment.	Hello all my BC Sisters! Just wanted to give some hope and encouragement to all of you ... Today, I "graduated" from the chemo phase ... and I feel liberated! ... I have completed 4 Dose Dense AC, plus 12 weekly Taxol.Sixteen treatments in all...
4	breast, get, know, biopsy, lump, wait	Post-test discussions.	Okay, as many of you know, I had my biopsy last Thursday. No word yet on results. I just don't get why the follow ups after these tests can't be with the radiologist who reads and understands these tests in the first place. Sigh! Any thoughts and advice would be much appreciated.
5	cancer, scan, radiation, bone, chemo, pain	Share diagnostic test results.	3 weeks ago my mom was diagnosed with moderatly invasave ductal carcinoma with necrosis. This was scary enough.
6	thank, love, prayer, hug, help, hope	Pray for others.	Dear Lord, please look down upon these beautiful, loving, sweet women…Please let my sisters ... feel your healing grace … Please protect them ...
7	breast, pain, surgery, feel, anyone, look	Anxiety after getting a treatment/test.	Today I got a bone scan, which of course brought an intense wave of scanxiety as soon as I walked into the hospital.
8	cancer, breast, chemo, node, stage, lymph	Sentinel nodes biopsy.	You face some difficult decisions. To your specific question, unfortunately it is difficult to do a sentinel node biopsy after a mastectomy has been done.
9	cancer, year, diagnosis, thank, find, doctor	Concerns about family/genetic issue.	My maternal grandmother had bc… My mother had widespread sites of DCIS and had a mastectomy… I've been working on a family "cancer map" in preparation for genetic testing.
10	know, post, think, test, decision, ask	Treatment options.	I understand and appreciate that ... why women choose to have a BMX. I'm not questioning that decision... the role of the doctor is to table all the options and explain the pros and cons of each options.

11	risk, study, treatment, cell, tamoxifen, patient	Clinical trials, research news, podcasts and study results.	There is a small advantage with an AI over tamoxifen in the first five years-in the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial …
12	hair, chemo, week, like, see, feel	Hair shedding (eyebrows) after chemo treatment.	All of my body hair fell out/off after my second chemo treatment with Cytoxan and Taxotere in early February - head hair, eybrows, arms, legs, genital region - entire body.
13	surgery, node, chemo, find, tumor, back,	Worries after surgeries (mostly lumpectomy).	I am worried about so many things. I am frightened and devastated. Any insight would be greatly appreciated! Thank you.
14	good, day, think, hope, work, well	Companionship and greetings.	Hi Jewels! Hope everyone is having a good Monday… Good day to stay inside and get things organized. Hope everyone has a great day!
15	get, people, life, friend, woman, say	Topics beyond health	In fact, in 2004 women made up 54 percent of the U.S. electorate, the highest percentage in history. Their interest in and impact on politics has been increasing.
16	pain, go, help, blood, http, name	Other resources, such as books, web sites, articles, centers.	I just read something posted by a doctor on another website that I found very interesting. Before you poo-poo this, please read it through. He seems to believe that …
17	le, arm, know, help, lymphedema, therapist	LE –lymphedema, which occurs after cancer surgery.	I was diagnosed with LE before I noticed any swelling, although I could sure feel the LE symptoms: tingling, heaviness, aching of my upper arm.
18	implant, breast, go, mom, exchange, surgery	Breast implants for breast reconstruction.	Beth, I'd be very interested if you have more information about your Beckers and what type of silicone was used. ... it appears that the Becker implant on the market 11 years ago would have been 75% saline with 25% soft silicone...
19	surgery, breast, one, time, bra, wear	Breast reconstruction (lumpectomy or mastectomy).	..., the difference between TRAM and DIEP is the veins they use to hook up the blood supply to the transplanted flap. With the DIEP… I went with the TRAM cause I had an experienced guy here doing those, but not DIEP.
20	life, chemo, year, good, work hope	Experiences of life changes after a breast cancer diagnosis.	I would just like to post a note about how my life has changed in just one year. I was diagnosed …I would wake up crying in the middle of the night wondering if I'd see my 9th grader graduate. Today, I am cautiously optimistic about my future.

Table 3. The performance of different predictive models with topics (*K=6, M=1*).

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
Logistic Regression+Topics	0.902	0.848	**0.889**	0.762
Decision Tree+Topics	0.759	0.781	0.832	0.685
Random Forest+Topics	0.891	0.832	0.874	0.747
GRU+Topics	0.903	**0.849**	**0.889**	0.762
P-LSTM+Topics	**0.904**	**0.849**	**0.889**	**0.765**

Table 4. Performance of predictive models with different number of topics (*K=6, M=1*).

Model	AUC			
The number of topics	K = 10	K = 20	K = 30	K = 40
Logistic Regression+Topics	0.902	0.902	0.903	0.903
Decision Tree+Topics	0.757	0.759	0.757	0.760
Random Forest+Topics	0.892	0.891	0.889	0.888
P-LSTM+Topics	**0.904**	**0.904**	**0.905**	**0.905**

3.2.5 Effects of feature selection.

To check the robustness of our P-LSTM+Topics model, we also apply two off-the-shelf dimensionality reduction methods -- Principle Component Analysis (PCA) and Random Forest feature ranking -- to check if the number of topics can be reduced before we feed them into our model. PCA is a data transformation method that linearly projects data into a variance-maximizing space from which a lower-dimensional subspace can be chosen [13,31]. Random Forest classification has also been applied as a method of feature selection [11,23], including two popular ways: permutation importance [1] and impurity importance [3]. We apply the second one in our experiment, which is derived from the training of a Random Forest classifier and selection of important features by Gini impurity [10,20]. We use Python package *sklearn* for feature selection. The experiments are still based on *K=6* and *M=1* with 10 times 10-fold cross validation. With PCA, we choose the number of components equals to 6, which can explain 99.99% variance for all features.

Results in Table 5 show that feature selections on topic features do not help to improve the predictive performance. The P-LSTM+Topics (shown in Table 3) model performs significantly better (with all p-values < 0.001 in paired t-tests) before applying the two feature selection methods.

Table 5. P-LSTM+Topics with different feature selection methods on topics (*K=6, M=1*).

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
P-LSTM+Topics with PCA feature selection	0.853	0.801	0.851	0.698
P-LSTM+Topics using Random Forest feature selection	**0.894**	**0.841**	**0.882**	**0.759**

3.2.6 Varying the observation period and prediction period.

To further check the robustness of our P-LSTM+Topics model, we vary the values of *K* and *M*. First, we increase the length of observation period to *K=12*. This leads to fewer instances in our dataset. After removing instances with 12 consecutive zeroes, we have 269,390 instances, with a class ratio of 3:1 (inactive vs active). The same set of baseline and proposed models are trained and evaluated using 10 times 10-fold cross validation. Results are shown in Table 6.

Table 6. The performance of predictive models (*K=12, M=1*).

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
Logistic Regression	0.909	0.872	0.920	0.669
Decision Tree	0.722	0.847	0.900	0.673
Random Forest	0.905	0.882	0.924	0.737
LSTM	0.917	0.887	0.927	0.746
P-LSTM+Topics	**0.919**	**0.888**	**0.928**	**0.745**

As one would expect, compared to *K=6* (shown in Table 1), a longer observation period leads to better predictive performance. The performance of proposed P-LSTM+Topics model is consistent: it still outperforms our baseline models (p-values<0.05 in paired t-test for AUCs), although the improvement of the Parallel LSTM with Topics over the simpler LSTM is not statistically significant (p-value=0.16).

Then we extend the prediction period to the next six months (i.e., *M=6*). The settings and procedures are the same as previous experiments, including removing instances with *K* consecutive zeroes. Results based on 10 times 10-fold cross validations are listed in Table 7 and Table 8 respectively.

Compared to predictions of the next one month (*M=1*), predicting posting activities during a longer future period is generally more challenging, evidenced by lower AUCs for both settings. In addition, the inactive class has lower F1 scores, while the active class actually has slightly higher F1 scores. This suggests that the challenge in predicting longer-term posting activities is mainly caused by difficulties in predicting longer hiatus during 6 months than predicting inactivity during one month.

Table 7. The performance of predictive models (*K=6, M=6*).

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
Logistic Regression	0.852	0.750	0.775	0.718
Decision Tree	0.792	0.757	0.767	0.745
Random Forest	0.851	0.776	0.778	0.773
LSTM	0.865	0.785	0.785	0.785
P-LSTM+Topics	**0.869**	**0.788**	**0.786**	**0.789**

Table 8. The performance of predictive models (*K=12, M=6*).

Model	AUC	Accuracy	F1 (inactive)	F1 (active)
Logistic Regression	0.865	0.781	0.840	0.654
Decision Tree	0.735	0.780	0.826	0.700
Random Forest	0.865	0.807	0.845	0.744
LSTM	0.878	0.815	0.853	0.750
P-LSTM+Topics	**0.881**	**0.818**	**0.856**	**0.754**

At the same time, the performance of P-LSTM+Topics is still consistent: they have better performance than baseline models (all p-values<0.001 in paired t-tests for AUCs). When *K=6, M=6*, the P-LSTM+Topics model performs slightly better than LSTM (all p-values<0.01 in paired t-tests for AUCs). With *K=12, M=6*, P-LSTM+Topics only performs marginally better than LSTM (p-value=0.058 in paired t-tests for AUCs).

3.3 The impact of discussion topics

Now we know that adding discussion topics from posts can help to improve the prediction of user posting activities. Inspired by inverse classification [19], we would like to explore which topics are more sensitive to possible interventions that aim at retaining users' posting activities. Specifically, if an OHC predicts a user to be inactive in the next month, but wants to keep her active in posting, which topics should the OHC recommend to the user so that when the user have more coverages on these topics, her prediction can switch to the other class (i.e., active)?

To investigate how changes to different discussion topics would affect predictions from our model, we first use a parallel LSTM predictive model to assign a class label for each instance. Then, for given a topic, we add 1 standard deviation to that topic's probability during the last month of observation (i.e., the 6th month if *K=6*) for all instances. Such modified instances with a new topic distribution for the last month then get new predicted labels from the same parallel LSTM predictive model. Comparing new labels and previous labels for each instance, we can find how many instances that were previous predicted as inactive switch to the active class in new predictions after increasing coverage on the artificially augmented topic. We run such experiments for each of the 20 topics. Figure 4 illustrates the process.

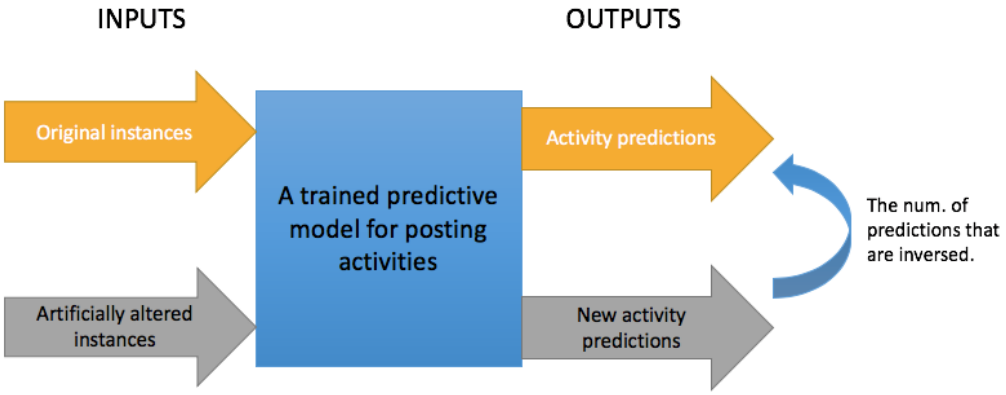

Figure 4: The procedure of our inverse classification experiments.

Table 9. Top topics based on the number of inversely classified instances with (*M=1*, and *K=6* and *12* respectively).

K = 6		K = 12	
Topic ID	Num. of Instances	Topic ID	Num. of Instances
14	**6726**	**14**	**5126**
3	**5548**	**3**	**4473**
6	**5066**	**6**	**4039**
19	3901	19	3107
10	3721	15	2935
15	3671	10	2747

Table 9 lists the top 6 topics by the number instances whose predictions have been "inversed" (from the inactive class to the active class). These top topics are consistent no matter whether we use predictive models trained with *K=6 M=1* or *K=12 M=1*. We focus on the top 3 topics (in bold) as their numbers of inversely classified instances are much higher than the other 3 topics. Topic 14 is about "Companionship and greetings"; Topic 3 is related to sharing personal experience and emotion with treatments; Topic 6 focuses mainly "Praying for others". In other words, users who discuss more on these topics during the last month of observation periods would have the best chance of having their predictions inversed from "inactive" to "active".

The results also concur with findings from empirical studies on social support. Companionship, also known as network support, is about informal chatting, humor, teasing and discussion of daily life. Emotional support is about empathy, encouragement, affection, caring etc. Research has previous revealed that OHC users' involvement in companionship [35] and emotional support [36] serve as an indicator of their long-term engagement in an OHC.

4 Conclusions and Future Work

This paper aims at predicting users' posting activities in OHCs. We propose a new framework to set up the prediction problem that is less arbitrary and more flexible than traditional user churn prediction for online communities and social media. In the framework, the length of the observation period and the prediction period can change based on the need of community management, and the observation period does not have to start from the first month of a user's online activities. This framework can thus provide predictions for any user whose temporal spans of online activities are longer than the observation window. It also allows predictions for short-term activities, long-term hiatus, or user churn. Our proposed approach leverages deep learning methods and user-generated content. It feeds one behavior trajectory in posting volume and another behavior trajectory in post topics into a parallel LSTM. The first trajectory is based on the number of posts over time and does not need feature engineering, while the second trajectory is based on topic distribution overtime, which is generated with unsupervised topic modeling and does not need human annotations. Through various experiments, we demonstrate that our P-LSTM+Topics model consistently outperforms baseline methods and that adding the trajectories on topics distributions helps to improve predictive performance.

Such predictions can help OHC managers improve user retention and community management via interventions such as email reminders or content recommendations. Thus, we also investigate what discussion topics are the most sensitive to possible interventions via inverse classification. It turns out discussions on companionship and emotional support top OHC mangers' list if they want to inverse more instances' predictions from being inactive to being active.

This research is not without limitations. We only work with data from one OHC, and test two different lengths for observation periods and prediction periods respectively. The predictive model is only for and based on posting activities. While users' clicking behaviors are also important, we do not have access to clickstream data for this OHC.

There are also several directions for further research. For example, we would also like to examine what users talk about before they suddenly disappear or when they re-appear after a long hiatus. Given the nature of cancer, we suspect that these are related to their offline health. Designing and evaluating interventions for user retention based on our predictions would be an interesting undertake. As some users are cancer survivors themselves while some are caregivers, and they come to an OHC with different purposes. Thus, being able to differentiate their roles can potentially help us better predict their future posting behaviors.

ACKNOWLEDGMENTS

Kang Zhao's work has been partially supported by the National Natural Science Foundation of China (Award #: 71572013).

REFERENCES

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (May 2010), 1340–1347. DOI:https://doi.org/10.1093/bioinformatics/btq134

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan (2003), 993–1022.

[3] Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* 143, (March 2020), 106839. DOI:https://doi.org/10.1016/j.csda.2019.106839

[4] Roger Burrows, Sarah Nettleton, Nicholas Pleace, Brian Loader, and Steven Muncer. 2000. VIRTUAL COMMUNITY CARE? SOCIAL POLICY AND THE EMERGENCE OF COMPUTER MEDIATED SOCIAL SUPPORT. *Information, Commun. Soc.* 3, 1 (January 2000), 95–121. DOI:https://doi.org/10.1080/136911800359446

[5] Langtao Chen. 2019. A Classification Framework for Online Social Support Using Deep Learning. In *International Conference on Human-Computer Interaction*. Springer, 178–188. DOI:https://doi.org/10.1007/978-3-030-22338-0_14

[6] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

[7] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A Nanavati, and Anupam Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology Advances in database technology - EDBT '08*, ACM Press, New York, New York, USA, 668. DOI:https://doi.org/10.1145/1353343.1353424

[8] Christine Dunkel-Schetter. 1984. Social Support and Cancer: Findings Based on Patient Interviews and Their Implications. *J. Soc. Issues* 40, 4 (January 1984), 77–98. DOI:https://doi.org/10.1111/j.1540-4560.1984.tb01108.x

[9] Susannah Fox. 2011. *The social life of health information, 2011*. Pew Internet & American Life Project Washington, DC.

[10] Hong Han, Xiaoling Guo, and Hua Yu. 2016. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th ieee international conference on software engineering and service science (icsess)*, IEEE, 219–224.

[11] Md Al Mehedi Hasan, Mohammed Nasser, Shamim Ahmad, and Khademul Islam Molla. 2016. Feature Selection for Intrusion Detection Using Random Forest. *J. Inf. Secur.* 07, 03 (2016), 129–140. DOI:https://doi.org/10.4236/jis.2016.73009

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (November 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735

[13] Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374, 2065 (April 2016), 20150202. DOI:https://doi.org/10.1098/rsta.2015.0202

[14] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2018. Identifying emotional support in online health communities. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[15] Amanda Leiter, Tomasz Sablinski, Michael Diefenbach, Marc Foster, Alex Greenberg, John Holland, William K Oh, and Matthew D Galsky. 2014. Use of Crowdsourcing for Cancer Clinical Trial Development. *JNCI J. Natl. Cancer Inst.* 106, 10 (October 2014). DOI:https://doi.org/10.1093/jnci/dju258

[16] Xi Long, Wenjing Yin, Le An, Haiying Ni, Lixian Huang, Qi Luo, and Yan Chen. 2012. Churn analysis of online social network users using data mining techniques. In *Proceedings of the international MultiConference of Engineers and Conputer Scientists*.

[17] Yingjie Lu. 2013. Automatic topic identification of health-related messages in online health community using text classification. *Springerplus* 2, 1 (December 2013), 309. DOI:https://doi.org/10.1186/2193-1801-2-309

[18] Diane Maloney-Krichmar and Jenny Preece. 2005. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Trans. Comput. Interact.* 12, 2 (June 2005), 201–232. DOI:https://doi.org/10.1145/1067860.1067864

[19] Michael V Mannino and Murlidhar V Koushik. 2000. The cost-minimizing inverse classification problem: a genetic algorithm approach. *Decis. Support Syst.* 29, 3 (October 2000), 283–300. DOI:https://doi.org/10.1016/S0167-9236(00)00077-4

[20] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10, 1 (2009), 213. DOI:https://doi.org/10.1186/1471-2105-10-213
[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv Prepr. arXiv1301.3781* (January 2013). Retrieved from http://arxiv.org/abs/1301.3781
[22] Blaise Ngonmang, Emmanuel Viennet, and Maurice Tchuente. 2012. Churn prediction in a real online social network using local community analysis. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, IEEE, 282–288.
[23] Jie Ning and Robert G Beiko. 2015. Phylogenetic approaches to microbial community classification. *Microbiome* 3, 1 (December 2015), 47. DOI:https://doi.org/10.1186/s40168-015-0114-5
[24] Jennifer L Pearson, Michael S Amato, George D Papandonatos, Kang Zhao, Bahar Erar, Xi Wang, Sarah Cha, Amy M Cohn, and Amanda L Graham. 2018. Exposure to positive peer sentiment about nicotine replacement therapy in an online smoking cessation community is associated with NRT use. *Addict. Behav.* 87, (December 2018), 39–45. DOI:https://doi.org/10.1016/j.addbeh.2018.06.022
[25] Kenneth Portier, Greta E Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, and J. Yen. 2013. Understanding Topics and Sentiment in an Online Cancer Survivor Community. *JNCI Monogr.* 2013, 47 (December 2013), 195–198. DOI:https://doi.org/10.1093/jncimonographs/lgt025
[26] Jagat Sastry Pudipeddi, Leman Akoglu, and Hanghang Tong. 2014. User churn in focused question answering sites. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, ACM Press, New York, New York, USA, 469–474. DOI:https://doi.org/10.1145/2567948.2576965
[27] Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E Greer, and Kenneth Portier. 2011. Get online support, feel better--sentiment analysis and dynamics in an online cancer survivor community. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, IEEE, 274–281.
[28] Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. 2018. A Study of WhatsApp Usage Patterns and Prediction Models without Message Content. *arXiv Prepr. arXiv1802.03393* (February 2018). Retrieved from http://arxiv.org/abs/1802.03393
[29] Matthew Rowe. 2013. Mining user lifecycles from online community platforms and their application to churn prediction. In *2013 IEEE 13th International Conference on Data Mining*, IEEE, 637–646.
[30] NehaShankar Sharma. 2015. Patient centric approach for clinical trials: Current trend and new opportunities. *Perspect. Clin. Res.* 6, 3 (2015), 134. DOI:https://doi.org/10.4103/2229-3485.159936
[31] Fengxi Song, Zhongwei Guo, and Dayong Mei. 2010. Feature selection using principal component analysis. In *2010 international conference on system science, engineering design and manufacturing informatization*, IEEE, 27–30.
[32] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. 2015. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, 4041–4049.
[33] Xi Wang, Kang Zhao, Sarah Cha, Michael S Amato, Amy M Cohn, Jennifer L Pearson, George D Papandonatos, and Amanda L Graham. 2019. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. *Decis. Support Syst.* 116, (January 2019), 26–34. DOI:https://doi.org/10.1016/j.dss.2018.10.005
[34] Xi Wang, Kang Zhao, and Nick Street. 2014. Social Support and User Engagement in Online Health Communities. In *International Conference on Smart Health*. Springer, 97–110. DOI:https://doi.org/10.1007/978-3-319-08416-9_10
[35] Xi Wang, Kang Zhao, and Nick Street. 2017. Analyzing and Predicting User Participations in Online Health Communities: A Social Support Perspective. *J. Med. Internet Res.* 19, 4 (April 2017), e130. DOI:https://doi.org/10.2196/jmir.6834
[36] Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave? In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, ACM Press, New York, New York, USA, 833. DOI:https://doi.org/10.1145/2145204.2145329
[37] Ronghua Xu and Qingpeng Zhang. 2016. Understanding Online Health Groups for Depression: Social Network and Linguistic Perspectives. *J. Med. Internet Res.* 18, 3 (March 2016), e63. DOI:https://doi.org/10.2196/jmir.5042
[38] Ronghua Xu, Jiaqi Zhou, Qingpeng Zhang, and James A Hendler. 2018. Research on Online Health Communities: A Systematic Review.
[39] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. 2018. I Know You'll Be Back. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, New York, NY, USA, 914–922. DOI:https://doi.org/10.1145/3219819.3219821
[40] Christopher C Yang and Ling Jiang. 2018. Enriching user experience in online health communities through thread recommendations and heterogeneous information network mining. *IEEE Trans. Comput. Soc. Syst.* 5, 4 (2018), 1049–1060.
[41] Haodong Yang and Christopher C Yang. 2015. Using Health-Consumer-Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis. *ACM Trans. Intell. Syst. Technol.* 6, 4 (July 2015), 1–27. DOI:https://doi.org/10.1145/2700482
[42] Mi Zhang and Christopher C Yang. 2015. Using content and network analysis to understand the social support exchange patterns and user behaviors of an online smoking cessation intervention program. *J. Assoc. Inf. Sci. Technol.* 66, 3 (March 2015), 564–575. DOI:https://doi.org/10.1002/asi.23189
[43] Shaodian Zhang, Edouard Grave, Elizabeth Sklar, and Noémie Elhadad. 2017. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *J. Biomed. Inform.* 69, (2017), 1–9.
[44] Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We Make Choices We Think are Going to Save Us. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, ACM Press, New York, New York, USA, 1073–1081. DOI:https://doi.org/10.1145/3041021.3055134
[45] Kang Zhao, Greta E Greer, John Yen, Prasenjit Mitra, and Kenneth Portier. 2015. Leader identification in an online health community for cancer survivors: a social network-based classification approach. *Inf. Syst. E-bus. Manag.* 13, 4 (November 2015), 629–645. DOI:https://doi.org/10.1007/s10257-014-0260-5
[46] Kang Zhao, Xi Wang, Sarah Cha, Amy M Cohn, George D Papandonatos, Michael S Amato, Jennifer L Pearson, and Amanda L Graham. 2016. A Multirelational Social Network Analysis of an Online Health Community for Smoking Cessation. *J. Med. Internet Res.* 18, 8 (August 2016), e233. DOI:https://doi.org/10.2196/jmir.5985
[47] Binjun Zhu, Xiaofeng Cai, and Ruichu Cai. 2018. Answer Quality Evaluation in Online Health Care Community. In *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, Atlantis Press, Paris, France. DOI:https://doi.org/10.2991/ncce-18.2018.143