Noname manuscript No.
(will be inserted by the editor)

Syndromic Surveillance of Flu on Twitter Using Weakly Supervised Temporal Topic Models

Liangzhe Chen · K. S. M. Tozammel Hossain · Patrick Butler · Naren Ramakrishnan · B. Aditya Prakash

Received: date / Accepted: date

Abstract Surveillance of epidemic outbreaks and spread from social media is an important tool for governments and public health authorities. Machine learning techniques for nowcasting the flu have made significant inroads into correlating social media trends to case counts and prevalence of epidemics in a population. There is a disconnect between data-driven methods for forecasting flu incidence and epidemiological models that adopt a state based understanding of transitions, that can lead to sub-optimal predictions. Furthermore, models for epidemiological activity and social activity like on Twitter predict different shapes and have important differences.

In this paper, we propose two temporal topic models (one unsupervised model as well as one improved weakly-supervised model) to capture hidden states of a user from his tweets and aggregate states in a geographical region for better estimation of trends. We show that our approaches help fill the gap between phenomenological methods for disease surveillance and epidemiological models. We validate our approaches by modeling the flu using Twitter in multiple countries of South America. We demonstrate that our models can consistently outperform plain vocabulary assessment in flu case-count predictions, and at the same time get better flu-peak predictions than competitors. We also show that our fine-grained modeling can reconcile some contrasting behaviors between epidemiological and social models.

Keywords Syndromic Surveillance · Social Media · Topic Model · Hidden Markov Model

E-mail: {liangzhe, tozammel, pabutler, naren, badityap}@cs.vt.edu
Tel.: +1-540-231-0906
Fax: +1-540-231-4240
Address: 114 McBryde Hall (0106)
Department of Computer Science
Virginia Tech.
Blacksburg, VA 24061, USA

1 Introduction

Web searches and social media sources, such as Twitter and Facebook, have emerged as surrogate data sources for monitoring and forecasting the rise of public health epidemics. The celebrated example of such surrogate sources is arguably Google Flu Trends where user query volume for a handcrafted vocabulary of keywords is harnessed to yield estimates of flu case counts. Such surrogates thus provide an easy-to-observe, indirect, approach to understanding population-level health events.

The recent research has brought intense scrutiny on Google Flu Trends, often negative. Lazer et al. (2014) provide explanations for Google Flu Trend's lackluster performance. Some of the reasons are institutional (e.g., a cloud of secrecy about which keywords are used in the model, affecting reproducibility and verification), some are operational (e.g., lack of periodic re-training), while others could be indicative of more systemic problems, e.g., that the vocabulary for tracking might evolve over time, or that greater care is needed to distinguish which aspects of search query volume should be used in modeling. These problems are not unique to Google Flu Trends; they would resurface with any syndromic surveillance strategy, e.g., developing a flu count modeler using Twitter.

Motivated by such considerations, we aim to better bridge the gap between syndromic surveillance strategies and contagion-based epidemiological modeling. such as SI, SIR, and SEIS (Anderson and May, 1991). In particular, while models of social activity have been inspired by epidemiological research, recent work (Matsubara et al., 2012; Yang and Leskovec, 2011; Romero et al., 2011) has shown that there are key aspects along which they differ from biological contagions. Specifically, evidence from Matsubara et al. (2012); Crane and Sornette (2008) shows that the activity profile (or the number new people using a hashtag/keyword) shows a power-law drop—in contrast standard epidemiological models exhibit an exponential drop (Hethcote, 2000). Also, there is some evidence that hashtags of different topics show an exposure curve which is not monotonic, resembling a complex contagion (Romero et al., 2011).

In this paper, we show that we can reconcile the apparently contrasting behaviors with a finer-grained modeling of biological phases as inferred from tweets. For example, sample tweets "Down with flu. Not going to school." and "Recovered from flu after 5 day, now going to the beach" denote different states of the users (also see Figure 1). We argue that correcting for which epidemiological state a user belongs, the social and biological activity timeseries are actually similar. Hashtags and keywords merge users belonging to different epidemiological phases. We separate these states by using a temporal topic model in our paper. In addition, thanks to the finer-grained modeling, our approach gets better predictions of the incidence of flu-cases than direct keyword counting and also sometimes gets better predictions of flu-peaks than sophisticated methods like Google Flu Trends.

Our contributions are:

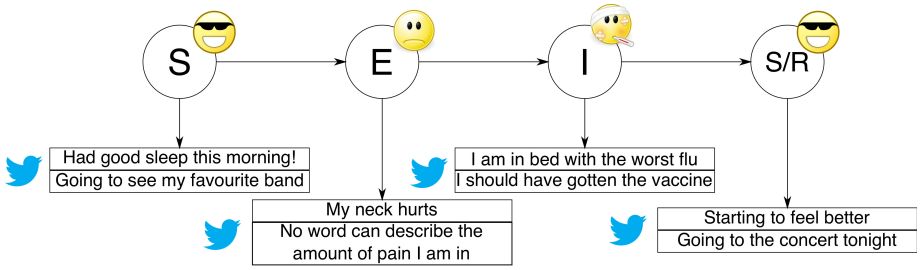


Fig. 1 A toy example showing possible user states and a tweet pattern associated with each state when a user is infected with flu for a time period

1. We propose temporal topic models (HFSTM and HFSTM-A) for inferring hidden biological states for users, and an EM-based learning algorithm for modeling the hidden epidemiological state of a user. The HFSTM-A model is robust to noisy and large vocabularies.
2. We show via extensive experiments using tweets from South America that our learners indeed learn meaningful word distributions and state transitions. Further, our methods can better forecast the flu-trend as well as flu-peaks by aggregating user states in a region over a time period.
3. Finally, we show that the state information learned by our models reconciles the social contagion activity profile with standard epidemiological models.

Our work can be seen as a stepping stone to better understanding of contagions that occur in both biological and social spheres. The rest of the paper is organized as follows: we review the related work in Section 2. We introduce our initial HFSTM model and its limitation, then we propose an improved model HFSTM-A to address this issue in Section 3. We describe our experiments in Section 4, and finally make conclusion and future work plan in Section 5 and Section 6.

2 Related Work

This is an extended work of our previous conference publication (Chen et al., 2014). In this study, we propose an improved model (HFSTM-A) to capture the latent health states of twitter users (Sec. 3.3). This model can handle documents with large and noisy vocabularies, which is not achievable with the initial HFSTM model. We show the inference algorithm for the new HFSTM-A model (Sec. 3.4). We further expand our test cases to include data from Argentina and Chile, and illustrate how HFSTM-A achieves as good results as HFSTM with a much larger and noisier vocabulary (Sec. 4). We also include the latest related work (Sec. 2), expand our future work (Sec. 6), and polish the overall writing.

2.1 Epidemiology

In the epidemiological domain, various compartmental models (which explicitly model states of each user) are employed to study the characteristics of flu diffusion (Hethcote, 2000). Some of the best known examples of such models are SI (Jacquez and Simon, 1993), SIR (Beretta and Takeuchi, 1995), and SEIS (Li and Muldowney, 1995), which are regularly used to model true flu case counts. Recently, several papers (Matsubara et al., 2012; Yang and Leskovec, 2011) show that the social activity profiles do not exactly follow these models, and propose several other variants. Note that different epidemiological models are used for different diseases, in this paper we focus our work on flu since it is very common disease.

2.2 Social Media

In the social media domain, related research has observed many strides in the last decade. Extensive data generated by these social networking sites (SNS) are being used to predict and forecast various societal events (Zhao et al., 2011), finding user interests (Spasojevic et al., 2014), or finding trending topics (Yang et al., 2014b). In particular the study of topic and word trends has become an important predictor for real world events and news. These trends are much easier and faster to get from social media than from traditional methods (e.g. reliable CDC case counts typically have lags of more than a month) (Glance et al., 2004). For disease prediction and forecasting, especially for flu, various methods have been proposed for large-scale (Ginsberg et al., 2008) and small-scale predictions (Christakis and Fowler, 2010). Furthermore, there are prediction methods that are solely based on Twitter (Lee et al., 2013; Culotta, 2010). Sadilek et al. (2012) and Brennan et al. (2013) studied the impact of different kinds of interactions to personal health–they calculate several features and predict the infection cases by classifications– in contrast, we directly model the overall state transitions for all users. Lamb et al. (2013) discriminate tweets that express awareness of the flu from those with actual infections, and train a classifier by which a user can tell if the author of a tweet is really infected. Aramaki et al. (2011) also trained classifiers for similar purposes. While their work is single-tweet-based, ours takes the tweet history into account. A tweet completely non-flu related is possible to be labeled as infected by our method if the tweets before and after both show signs of infection. Achrekar et al. (2011), Culotta (2010), and Lampos et al. (2010) fit a flu trend by analysing tweets via various methods including keyword analysis, and compare their flu trend fitting with CDC results. Lampos et al. (2010) present an automated tool using keywords to track the prevalence of Influenza-like Illness (ILI). These methods are very coarse-grained—they do not provide understanding on how the health state of a user changes over time, while we link the change of tweet pattern with standard epidemiological models. The unpublished recent work by Li and Cardie (2013) builds a Markov network to

capture the spatio and temporal relations between different locations. Their definition of states is based on the number of infections in a location (such as rising state, declining state), but states in our work are *epidemiological* states and they are learned directly from the tweet corpus.

2.3 Topic Model

In this paper, we use a variation of topic models for our purposes. The earliest topic modeling using LDA (Latent Dirichlet Allocation) (Blei et al., 2003) gained popularity for modeling a large amount of text documents (see (Blei et al., 2010) for review). Many variations of LDA have been proposed to model various problems. For modeling health related topics Paul et al. proposed the Ailment Topic Aspect Model (ATAM+) (Paul and Dredze, 2011) to capture various ailments from a corpus of tweets. This model is based on a topic aspect model (Paul and Girju, 2010), author-topic model (Steyvers et al., 2004), and it does not consider the temporal information of the text messages (as we do in this paper). Another variant of LDA is temporal topic models which can be categorized into two groups: Markovian and non-Markov. Wang and McCallum (2006) propose a non-Markov continuous time model for topic trends which can not be used to predict the user states. Gruber et al. propose a hidden topic Markov model (HTMM) (Gruber et al., 2007), which assumes that all the words in a sentence have the same topic and there may be a topic transition between two consecutive sentences. In the paper (Andrews and Vigliocco, 2010), Andrews et al. proposes a hidden Markov topic model (HMTM) that assumes that there is a topic transition between two consecutive words within a document. In the paper (Blasiak and Rangwala, 2011), Blasiak et al. uses a hidden Markov model to capture topic transition within documents which are subsequently used to classify new messages. These methods only capture transition of topics within a document or a message, they do not capture state transition of users *across* tweets. There are two other variants of LDA (Blei and Lafferty, 2006; Hong et al., 2011) studying the evolution of topic distributions over time, while our model studies the transition between a set of topic distributions which does not evolve over time. Moreover, their models do not capture the topic changes between consecutive messages of a user. Another recent related work is by Yang et al. (2014a) who combine keyword distributions with a shortest path algorithm to find out a monotonically increasing stage progression of an event sequence. In our problem, flu states are not monotonic, and have transition probabilities, which their method does not learn.

3 Formulation of Models

We formulate our models in this section. The hypothesis is that a tweet stream generated by a user can be used to capture the underlying health condition of

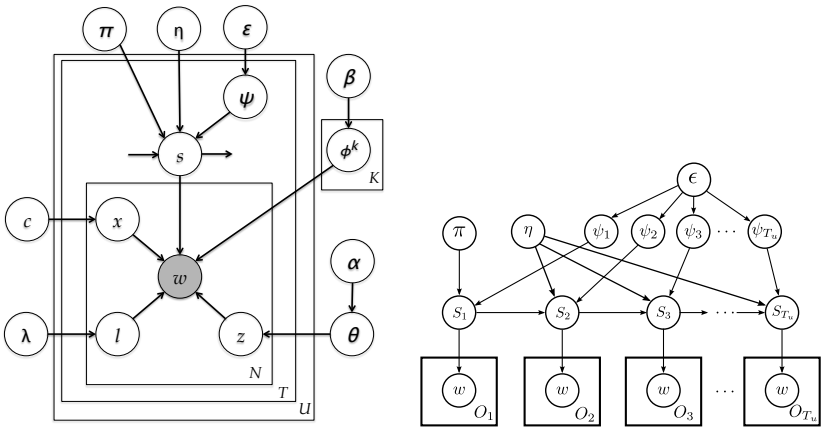
Table 1 Symbols used for HFSTM and HFSTM-A

Symbol	Meaning
S	Flu state
S_t	Flu state for the t-th tweet
ϵ	State switching parameter
π	Initial state distribution
η	Transition probability matrix
l	Binary background switching variable
x	Binary switch between flu and non-flu words
y	Aspect of word
λ	Parameter for the Bernoulli distribution for l
c	Parameter for the Bernoulli distribution for x
ϕ	Topic distribution
σ	Prior for l when aspect is introduced
γ	Prior for x when aspect is introduced
T_u	Total number of tweets for the u-th user
N_t	Total number of words in t-th tweet
w	Word variable in the template model
w_{tn}	the n-th word in the t-th tweet
z	Non-flu related topic
θ	Prior for non-flu related topics
α	Hyper parameter for topic distributions
ψ	State switching variable
K	Total number of states
β	Dirichlet parameter for word distributions
U	Number of users

that particular user; and that the health state (e.g., flu state) of a user remains the same within a tweet. Then we use our models to capture the flu states of a user which are S (healthy), E (exposed), or I (infected) based on the classic flu-like Susceptible-Exposed-Infected-Susceptible SEIS epidemiological model. These states model the different health conditions of a person throughout the lifecycle of the infection. In this study, we first introduce the HFSTM model. Then we show the limitation of HFSTM, and propose an improved model HFSTM-A (HFSTM with aspects) to address the issue.

3.1 Hidden Flu-State from Tweet Model (HFSTM)

A tweet is a collection of words and a tweet stream is a collection of tweets. The number of tweets varies across users and the number of words in a tweet varies within and across users. We denote the t-th tweet of a user by $O_t = \langle w_{t1}, w_{t2}, \ldots, w_{tN_t} \rangle$ where w_{tn} denotes the n-th word in the tweet and N_t denotes the total number of words in the tweet. Let $\mathcal{O}_u = \langle O_1, O_2, \ldots, O_{T_u} \rangle$ be the tweet stream generated by a user u and $\mathcal{S}_u = \langle S_1, S_2, \ldots, S_{T_u} \rangle$ be the underlying state of the stream $\mathcal{O}_u$. Here T_u denotes the length of the stream of a user u and $S_t \in \{S, E, I\}$. Let $\mathcal{O} = \langle \mathcal{O}_1, \mathcal{O}2, \ldots, \mathcal{O}_U \rangle$ be the collection of tweets for U users, from which we aim to learn the parameters of our model. We use K to denote the number of states that S_t can take (see Table 1 for notations).



(a) HFSTM (b) State transition

Fig. 2 (a) Plate notation for HFSTM: The variable S captures the hidden state of the user in which the user generated this tweet. N, T, U are the number of words, tweets, and users respectively. The LDA-like topic variable z capture non-flu related words. (b) HFSTM state variables expanded: Each message O_t is associated with a state S_t, which remains same for flu-related words in O_t. Switching from one state to another is controlled by a binary switching variable ψ and the next state S_{t+1} from the current state S_t is drawn using transition probabilities η

Our initial model—Hidden Flu-State from Tweet model HFSTM—is a probabilistic graphical model which captures the tweet structure of a flu-related tweet. It is a temporal topic model for predicting the state sequence of a user given $\mathcal{O}_u$ and is illustrated in Fig. 2(a). An expansion of the plate notation for the same is illustrated in Fig. 2(b). In this model each word w for $O_t \in \mathcal{O}_u$ is generated when the user is in a particular flu state (S_t) or the user talks about a non-flu related topic (z_i). For example, in the message "I have caught the flu. Feeling feverish. Not going to school" the words 'flu', 'feverish', 'caught' are generated because the user is in the "infected" state and the words 'going' and 'school' are generated by non-flu related topics. Sometimes a word can be generated due to noise which is also accounted for in our model.

The generative process for the model is shown in Alg. 1. A binary variable l determines whether or not a word is generated from a background distribution. The binary variable x determines whether the current word is generated from non-flu related topics or flu-state distributions. The value of l and x are generated from Bernoulli distributions parameterized by λ and c. The non-flu related topics follow the LDA like mechanism (Blei et al., 2003). The state for the first tweet is drawn from the initial distribution denoted by π. We assume that the states of the subsequent tweets are generated due to a state transition or by copying from the previous state which is determined by a binary switching variable ψ with prior parameter ϵ. The state S_t (for $2 \leq t \leq T_u$) of the subsequent tweets are drawn from transition matrix η and previous state S_{t-1} with probability ϵ or copied from the previous state S_{t-1} with probability

Algorithm 1 Generator($\lambda, c, \eta, \pi, \alpha, \beta, \epsilon$) for HFSTM

Input: A set of parameters.
Output: Topics and flu state of each user.

Set the background switching binomial λ
Choose $\phi \sim \text{Dir}(\beta)$ for the non-flu topics, flu states, and background distribution
Choose initial state $s_1 \sim \text{Mult}(\pi)$
Draw each row of η using $\text{Dir}(\alpha)$ ▷ Trans. matrix
Draw $\theta \sim \text{Dir}(\alpha)$
for each tweet $1 \leq t \leq T_u$ **do**
    if not the 1st tweet in the corpus **then**
        Draw $\psi_t \sim \text{Ber}(\epsilon)$
        if $\psi_t = 0$ **then**
            $S_t \leftarrow S_{t-1}$
        else
            $S_t \leftarrow \text{Mult}(\eta_{S_{t-1}})$
    for Each word w_i, $1 \leq i \leq N_t$ **do**
        Draw $l_i \in \{0, 1\} \sim \text{Ber}(\lambda)$ ▷ Background switcher.
        if $l_i = 0$ **then**
            Draw $w_i \sim \text{Mult}(\phi^B)$ ▷ Draw from background distribution.
        else
            Draw $x_i \in \{0, 1\} \sim \text{Ber}(c)$
            if $x_i = 0$ **then**
                Draw $z_i \sim \text{Mult}(\theta)$
                Draw $w_i \sim \text{Mult}(\phi^{z_i})$ ▷ Draw from non-flu related distribution.
            else
                Draw $w_i \sim \text{Mult}(\phi^{s_t})$ ▷ Draw from flu related distribution.

$1-\epsilon$. Once the state of a tweet is determined, a word is generated from a word distribution defined by that state.

Let $O_t = (w_1, \ldots, w_N)$ be the words that are generated when a user is in a particular state. The likelihood of the words generated by a user in that state is given below.

$$
\begin{aligned}
p(\mathcal{O}_u) &= \sum_{S_t} p(\mathcal{O}_u, S_t) = \sum_{S_t} p(O_1 \ldots, O_T, S_t) \\
&= \sum_{S_t} \sum_{S_{t-1}} p(O_t|S_t) p(S_t|S_{t-1)} p(O_{t-1}, S_{t-1}) \qquad (1)
\end{aligned}
$$

Andrews and Vigliocco (2010) show that such kind of likelihood function is intractable. In HFSTM the unknown parameters that we want to learn are $H = \{\epsilon, \pi, \eta, \phi, \lambda, c\}$. The posterior distributions over these unknown variables are also intractable since the posterior distributuions depend on the likelihood function. We hence developed an EM-based algorithm HFSTM-FIT to estimate the parameters H of the model (we omit the details for this algorithm, but rather elaborate on the inference algorithm for the extended HFSTM-A model in Sec. 3.4).

3.2 Issues with HFSTM

HFSTM requires a 'clean' vocabulary, i.e. a vocabulary that does not contain many background words. In real datasets, there is a huge imbalance between background and flu-related words. For example, among 100 tweets from a user, only two or three maybe related to his/her health state. As there is no supervision used in HFSTM, each word has the same probability of passing/failing the switches (see the parameters λ, c in 3.1), which biases our model towards background words. Hence it is likely for HFSTM to learn the complex state transitions among background words rather than among the flu-related words. If the dataset contains many tweets about some hot events such as a football game, the model would learn the state transition in the sport game rather than in the flu infection since the number of sport-related tweets overwhelms that of the flu-related tweets. For this reason, HFSTM needs a vocabulary that does not contain many background words—as a consequence, it highly depends on the accuracy of the selection of words, which decreases its generality.

3.3 Improving the model—HFSTM-A

Due to the issues with HFSTM, we propose a new model HFSTM-A (HFSTM with aspects) so that we can provide it with a larger noisier vocabulary. Our key idea is to explicitly include our belief of which words are likely to be useful for state transitions. Hence we add such weak supervision to HFSTM by introducing an 'aspect' value (y) for each word (a related approach has been used by Paul and Dredze (2011)). We call this new model HFSTM-A. This aspect y takes two values $\{0, 1\}$ based on whether the word is flu related or not. It then biases the switching probabilities so that background words are less likely to be explained by the state topic distributions. Note that this supervision is weak because the aspect of a word does not directly decide if a word is flu-related, it only increases/decreases the probability of a word being regarded as flu-related or not. Those words which we do not mark as related are still possible to be analysed by state topic distributions. As a result of the changes, HFSTM-A can handle much noisier vocabularies, at the mean time have comparable performance with the HFSTM model.

More concretely, in the plate notation of this new model HFSTM-A (see Fig. 3), y is the observed aspect value for a word, where l and x are the binary values which decide whether the word is generated by background topic, non-flu topics, or state topic distributions. In HFSTM, these two values are generated by the Bernoulli distribution with parameters λ and c. Now in HFSTM-A, y biases these probabilities and may thus change the values of l and x.

The generative process for the HFSTM-A model is shown in Alg. 2. In contrast to Alg. 1, we see that in Alg. 2 the value of l and x are now generated from Bernoulli distributions parameterized by λ_{y_i} and c_{y_i}, which are biased by

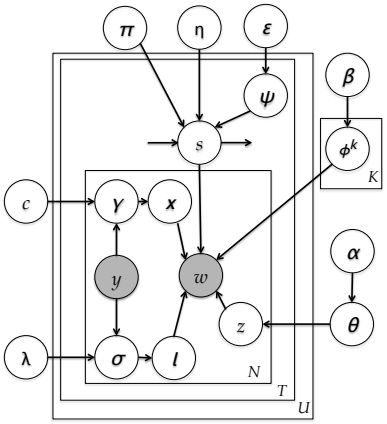


Fig. 3 Plate notation for HFSTM-A: The aspect value y is an observed variable for each word, and this variable is biases the probability of a word being generated by the various topics (See Sec. 3.3)

Algorithm 2 Generator$(\lambda, c, \eta, \pi, \alpha, \beta, \epsilon)$ for HFSTM-A

Input: A set of parameters.
Output: Topics and flu state of each user.

Set the background switching binomial λ
Choose $\phi \sim (\beta)$ for the non-flu topics, flu states, and background distribution
Choose initial state $s_1 \sim \text{Mult}(\pi)$
Draw each row of η using $\text{Dir}(\alpha)$ ▷ Trans. matrix
Draw $\theta \sim \text{Dir}(\alpha)$
for each tweet $1 \leq t \leq T_u$ **do**
    if not the 1st tweet in the corpus **then**
        Draw $\psi_t \sim \text{Ber}(\epsilon)$
        if $\psi_t = 0$ **then**
            $S_t \leftarrow S_{t-1}$
        else
            $S_t \leftarrow \text{Mult}(\eta_{S_{t-1}})$
    for Each word w_i, $1 \leq i \leq N_t$ **do**
        Draw $y_i \in \{0, 1\}$ (observed)
        Draw $l_i \in \{0, 1\} \sim \text{Ber}(\lambda_{y_i})$ ▷ Background switcher.
        if $l_i = 0$ **then**
            Draw $w_i \sim \text{Mult}(\phi^B)$ ▷ Draw from background distribution.
        else
            Draw $x_i \in \{0, 1\} \sim \text{Ber}(c_{y_i})$
            if $x_i = 0$ **then**
                Draw $z_i \sim \text{Mult}(\theta)$
                Draw $w_i \sim \text{Mult}(\phi^{z_i})$ ▷ Draw from non-flu related distribution.
            else
                Draw $w_i \sim \text{Mult}(\phi^{s_t})$ ▷ Draw from flu related distribution.

the observed aspect value y_i. The definition of λ_{y_i} and c_{y_i} are shown below.

$$\lambda_{y_i=0} = \lambda$$
$$\lambda_{y_i=1} = \lambda + b * (1 - \lambda)$$
$$c_{y_i=0} = c - a * c$$
$$c_{y_i=1} = c + a * (1 - c)$$

where a, b are the fixed biases we add to the switching probabilities. Basically, if a word is labeled as flu-related ($y_i = 1$), we increase its probability of passing the background switch ($\lambda_{y_i=1}$) and its probability of passing the non-flu topic switch ($c_{y_i=1}$), by pushing these probabilities closer to 1. In the equations above, we take a proportion (b and a respectively) of the residuals and add it to the probability; and if a word is not flu-related ($y_i = 0$), we decrease its probability of passing the non-flu topic switch ($c_{y_i=0}$), by pushing the probability towards 0 (we use a to shrink the value in the corresponding equation). Note that if a word is not flu-related, it can still be generated by non-flu topics. Hence its probability of passing the background switch ($\lambda_{y_i=0}$) is kept unbiased. In our experiments, we use $a = 0.4$, $b = 0.4$, and find the performance good and stable around these values.

3.4 HFSTM-A-FIT: Inference and Parameter Estimation

We next show an EM-based algorithm HFSTM-A-FIT to estimate the parameters $H = (\epsilon, \pi, \eta, \phi, \lambda, c)$ of our model.

At each time point t a user can be in any of the $2K$ states where the first K states denote that the user happens to be in the state due to a state transition from his state at time $t-1$ and the rest of states from $K+1 \dots 2K$ denote that the state of the user is simply copied from the state of the user at time $t-1$.

3.4.1 E-Step

We use forward-backward procedure for estimating parameters. We define the forward probability $A_t(i)$ and the backward probability $B_t(i)$ for a tweet stream as follows.

$$A_t(i) = P(O_1, O_2, \dots, O_t, S_t = i)$$
$$B_t(i) = P(O_{t+1}, \dots, O_{T_u} | S_t = i)$$

$A_t(i)$ is the joint probability of the partially observed sequence until time t and state S_t is i. $B_t(i)$ is the joint probability of the partially observed sequence from $t+1$ to T_u, given state S_t is i. Both $A_t(i)$ and $B_t(i)$ can be solved inductively. See the linked equations for more details on how $A_t(i)$ and $B_t(i)$ are calculated.

Let $\gamma_t(i)$ be the probability of being in state S_i at t_{th} tweet given the observed tweet sequence $\mathcal{O}_u$.

$$\begin{aligned}\gamma_t(i) &= P(S_t = i | \mathcal{O}_u) \\ &= \frac{A_t(i) B_t(i)}{\sum_{i=1}^{2K} A_t(i) B_t(i)}\end{aligned}$$

To estimate transition probability we define $\xi_t(i,j)$, the probability of being in state i at t and in state j at $t-1$ given the $\mathcal{O}_u$.

$$\begin{aligned}\xi_t(i,j) &= P(S_t = i, S_{t+1} = j|\mathcal{O}_u) \\ &= \frac{P(S_t = i, S_{t+1} = j, \mathcal{O}_u)}{P(\mathcal{O}_u)}\end{aligned}$$

3.4.2 M-Step

In this step we re-estimate the parameters ϵ, π, η, ϕ, c, λ. Due to space constraints only the estimations of π and η are shown as follows. Please check the Appendix for all the equations.

$$\pi_i = \frac{\sum_{u=1}^{U} \gamma_1(i)}{\sum_{u=1}^{U}\sum_{i=1}^{K} \gamma_1(i)}$$

$$\eta_{ij} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T} (\xi_t(i,j) + \xi_t(i+K,j))}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{j=1}^{K} (\xi_t(i,j) + \xi_t(i+K,j))}$$

4 Experiments

We describe our experimental results next. All the experiments are designed to answer the following questions:

1. Can HFSTM and HFSTM-A robustly learn in presence of different noise levels in a dataset? (see Sec. 4.2)
2. What are the state-topic distributions learnt by our models? (see Sec. 4.3)
3. Is the state transition table learned reasonable? (see Sec. 4.4)
4. How do our models perform for flu case-count and peak predictions? (see Sec. 4.5)
5. Finally, as mentioned in Sec. 2, several papers (Matsubara et al., 2012; Yang and Leskovec, 2011) have shown that the rising and falling pattern of keywords count in social media does not match with that in epidemiological model. By including the extra state information, can we bridge this gap between social and epidemiological activity? (see Sec. 4.6)

4.1 Experimental Set-up

First we describe our set-up in more detail. Our algorithms were implemented in Python.[1]

[1] Code and vocabulary can be found here: `http://people.cs.vt.edu/liangzhe/code/hfstm-a.html`

4.1.1 Vocabularies

To ensure that the most important words (directly flu-related words like 'flu', 'cold', 'congestion', etc.) are included in our vocabulary, we first build a flu-related keyword list. Chakraborty et al. (2014) construct a flu-keyword list, by first manually setting a seed set, then using two methods (pseudo-query and correlation analysis, see their paper for more details) to expand this seed set, and then finally pruning it to a 114 words keyword list. A similar keyword-construction procedure (expanding by crawling websites) was also used by Lampos and Cristianini (2012). For our experiments, we include the same 114 keywords from Chakraborty et al. (2014) first. We then include 116 words selected by our in-house experts, which are not directly related to flu, but may implicitly imply the state of a user, such as 'hopeless', 'bed', 'die', 'sad', etc. We use these (a total of 230) words as the vocabulary for HFSTM since it cannot deal well with a noisier vocabulary (see Sec. 3.2).

For HFSTM-A, the extension of HFSTM which is designed to handle larger vocabularies with much background noise, we enlarge the size of the vocabulary by simply adding the most frequent words in the corpus. After automatically extracting these top words, we get a final vocabulary of 2739 words.[2] All other words not in our vocabulary but occurring in the corpus are mapped to a single generic block-word. We label a word as 1 (flu-related) if it is in the previous 230 words list (note again this is only weak supervision, this label does not directly decide whether this word is generated by background topics, or state topics). Thanks to our model design, as we describe later, HFSTM-A is able to learn meaningful state transitions and topic distributions, inspite of having a more than $10X$ larger vocabulary.[1]

4.1.2 Datasets

We collected tweets generated from 15 countries in South America for the period Dec, 2012—Aug, 2014 using Datasift's Twitter collection service.[2] The Datasift twitter feed is enriched in two ways which are relevant to collecting the twitter flu data. The first is using the Basis technology[3] natural language processing facilities from which we use the lemmatized form of words to improve word count metrics. The second is a custom set of geocoding algorithms used to detect the location of a tweet since only 5% of tweets are actually geotagged. We then improve the quality of our dataset by removing bots, spammers, and retweets.

We create a training dataset *TrainData*, using the tweets from Jun 20, 2013 to Aug 06, 2013, which contains a peak of infections. We created three evaluation sets using tweets from different time-periods: *TestPeriod-1* (Dec 01, 2012 to Jul 08, 2013), *TestPeriod-2* (Nov 10, 2013 to Jan 26, 2014), and *TestPeriod-3* (Mar 01, 2014 to Aug 31, 2014). *TestPeriod-1* and *TestPeriod-2* are time periods before and after our training period in the same year (2013); We further

[2] `http://datasift.com/`

[3] `http://www.basistech.com`

test our models trained from 2013 on *TestPeriod-3*, which covers a complete flu season in the next year 2014. The number of flu related tweets (containing at least one flu keyword) for these test periods are $\sim 1.8M$, $\sim 0.3M$, and $\sim 4M$ respectively. For the two individual countries used in Sec. 4.5 for *TestPeriod-1*, this number is 60k for Chile, and 112k for Argentina. We use tweets that occurred during the flu season in 2013 as the training set for maximizing the number of samples that are tagged as infected. We choose the three test sets as they either contain a complete flu season, or contain interesting rising patterns (detecting the rising part of the disease is one of the most challenging tasks in surveillance (Butler, 2013)). For creating training data we perform keyword and phrase checking (from our vocabulary) to identify a set of users who have potentially tweeted a flu-related tweet. We then fetch their tweet streams from Twitter API for the training period. We then use the Datasift service to preprocessing these tweets (stemming, lemmatization, etc.), and get our final training dataset of roughly 34,000 tweets. Under such a setting, our inference algorithm HFSTM-A-FIT takes around 2 hours to run on a 4 Xeon E7-4850 CPU with 512GB of 1066Mhz main memory.

We collected data from The Pan American Health Organization (PAHO, 2012) for the ground-truth reference dataset for flu case counts (trends). PAHO is the ground-truth medical report source for South America and it plays the same role in South America as CDC does in the USA (CDC does not provide flu trend data for South America). Note that PAHO gives only per-week counts.

4.2 Robustness and Consistency (Q.1)

To first check the performance of our models under different conditions, we set up three kinds of simple synthetic datasets for the learners. We first choose a set of fixed parameters as base settings for generating a dataset. We then vary the background switching parameter (λ) for creating a set of datasets with different noise levels (to clarify, note that via λ we are only varying the number of background words in the dataset here, not in the vocabulary). For the third variant of datasets, we vary the number of users for generating a set of datasets. Firstly, in all the datasets, our learner was able to recover the true parameters, and show a good estimation of switching variables, transition probabilities and word distributions on these synthetic datasets. Table 2 shows the estimation error of π, η and the word distribution for each state, measured by the KL distance between the true parameter and the estimated value. Secondly we see that the performance of our models is pretty robust: it does not degrade with an substantial increase in noise level, and the learner is also stable when we increase the number of users. Finally, note that HFSTM-A learns similar quality results like HFSTM, inspite of a much enlarged vocabulary.

Expt	KL of π		KL of η		KL of ϕ_0		KL of ϕ_1		KL of ϕ_2		KL of ϕ_3	
	m1	m2	m1	m2	m1	m2	m1	m2	m1	m2	m1	m2
base	0.04	0.04	0.08	0.02	0.24	0.57	0.2	0.08	0.2	0.12	0.2	0.04
$\lambda = 0.1$	0.04	0.04	0.08	0.03	0.01	0.48	0.01	0.17	0.01	0.13	0.01	0.04
$\lambda = 0.3$	0.04	0.04	0.03	0.50	0.00	0.14	0.01	0.01	0.01	0.01	0.01	0.01
$\lambda = 0.5$	0.04	0.04	0.03	0.05	0.01	0.70	0.01	0.17	0.01	0.16	0.01	0.06
$\lambda = 0.9$	0.04	0.04	0.04	0.01	0.00	0.45	0.01	0.07	0.01	0.08	0.01	0.02
$U = 50$	0.04	0.04	0.29	0.26	0.04	0.27	0.07	0.01	0.06	0.01	0.09	0.01
$U = 70$	0.04	0.04	0.30	0.42	0.05	0.14	0.08	0.03	0.04	0.03	0.09	0.03
$U = 90$	0.04	0.04	0.08	0.01	0.02	0.52	0.03	0.07	0.03	0.02	0.03	0.01
$U = 110$	0.04	0.04	0.01	0.40	0.00	0.17	0.01	0.01	0.01	0.01	0.01	0.01
$U = 130$	0.04	0.04	0.00	0.01	0.00	0.71	0.01	0.04	0.01	0.04	0.01	0.01
$U = 150$	0.04	0.04	0.06	0.07	0.00	0.20	0.01	0.01	0.01	0.01	0.01	0.01

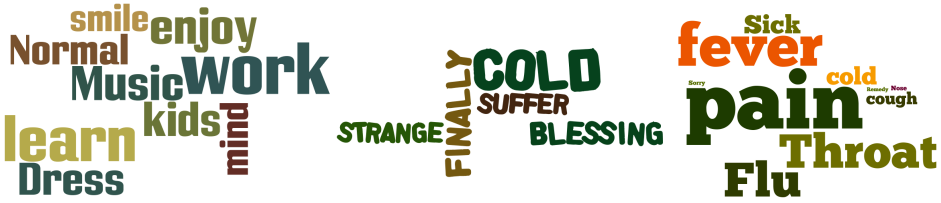
Table 2 Robustness and Consistency of our models (m1 = HFSTM, and m2 = HFSTM-A) using synthetic datasets. In the 'base' setting, we use 100 users, and a vocabulary of size 92, where the number of background words, state words, and non-flu topic words are 20, 60, and 12 respectively. We vary the the number of background words (by varying λ) and the number of user from 50 to 150. It can be seen that the performance of both models do not suffer from increasing noise levels in the dataset (different from the noise in the vocabulary), and it is pretty stable when we increase the number of users in the experiments

4.3 Word distributions learnt for each state (Q.2)

In short, our model learns meaningful topic word distributions for the flu states from real data (*TrainData*). See Figure 4—it shows a word cloud for each state-topic distribution (we renormalized each word distribution after removing the generic block-word) we learnt using HFSTM-A. Note that both HFSTM-A and HFSTM learn meaningful distributions, here we only show results from HFSTM-A since the result from HFSTM is similar. The most frequent words in each state matches well with the S(usceptible), E(xposed) and I(nfected) states in epidemiology. These word distributions in Fig. 4 correspond to the S, E, I states shown in Fig. 5. As shown in the figure, the S state has normal words, the E state starts to gather words which are indicating an exposure or approaching to the disease (and contains both 'S-like' and 'I-like' words), while the I state gets many typical flu-related words. The I state captures flu-related keywords like flu, fever, pain; the E state contains words like cold, suffer, strange; and the S state has words like enjoy, work, music, smile.

4.4 Transition probabilities learnt between states (Q.3)

We show the state transition diagram *learned* from real data (*TrainData*) by HFSTM-A in Figure 5. Again, HFSTM-A is as good as HFSTM, with a much larger vocabulary. The initial state probability learned is $[0.91, 0.02, 0.07]$, with high probability that a tweet starts at state S, and with much lower probabilities it starts at state E or I. We observe that for each state, it firstly has the tendency to stay in that state, which is reasonable because a twitter user is likely to post more than one tweet in any given state. When there



(a) S state (b) E state (c) I state

Fig. 4 The translated word cloud for the most probable words in the S, E and I state-topic distributions as learnt by HFSTM-A on *TrainData*. Words are originally learned and inferred in Spanish, we then translate the result using google translate for the ease of understanding. The size of the word is proportional to its probability in the corresponding topic distribution. Our model is able to tease out the differences in the word distributions between them

is a transition occuring, we see that transition between S and E is larger than between E and I, showing the fact that the probability of truly getting infected is lower than the probability of getting just exposed. Interestingly, these transition probabilities match closely with the standard epidemiological SEIS model and intuition.

We also investigate the most-likely state sequence for each user learned by HFSTM-A. Using the parameters learned by our model, we take a sequence of tweets from one user, and use MLE to estimate the state each tweet is in. Table 3 shows multiple examples of these transitions (we show the translated English version here using Google Translate and further refined by a native speaker) using HFSTM-A (the results are similar to that of HFSTM). As we can see, our model is powerful enough to learn the Exposed state, before the user is infected. This also shows the accuracy of our transition probabilities between the flu states.

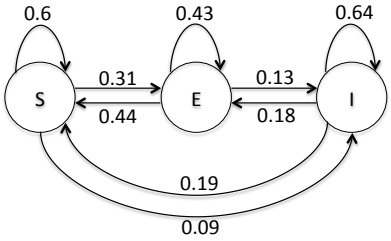


Fig. 5 The transition diagram between flu-states automatically learned by HFSTM-A.The probabilities are rounded up for simplicity. Note that the structure of the state transitions is close to the standard epidemiological SEIS model

User	Date	Tweet Message	State
1	4 Jul 2013	S: @ PauFigueroaentoces yes .... but by then I wouldn't like to feel like I feel now because I wouldn't be able to enjoy the vacations.	Healthy
	4 Jul 2013	I finished my first job, one more to go, and me feeling so bad... I want to rest.	Healthy
	4 Jul 2013	@ Kimy2Ramos My queen, I hope you're having a great time... because I feel terrible. I have a headache and fever =(... I love you a lot.	Exposed
	4 Jul 2013	@ PauFigueroaflu, with the flu, headache, body ache, and even my sight hurts... Couldn't ask for anything else.	Infected
	4 Jul 2013	time to studyyy...	Healthy
2	10 Jul 2013	I'm feeling like a boss for working on this by myself. I'm gonna pass, no doubt about it hahahaha.	Healthy
	10 Jul 2013	already Wednesday? Today to Aliados, how awesome.	Healthy
	11 Jul 2013	Any season is spring for me if I'm with you.	Exposed
	11 Jul 2013	It's just great how I got sick. Sad part is that I can't even miss school -.-	Exposed
	11 Jul 2013	It was so great to see a scene from Peter y Pablo, how much I missed those things.	Exposed
	11 Jul 2013	Oh, how much I hate you Tabcin. You're gross -. -	Healthy
	11 Jul 2013	Lately I've been missing those little things that made you so unique. I wonder where all those virtues went: S	Exposed
	11 Jul 2013	I'm feeling awful: fever, headache, dizziness, chest pain, snot, snot, snot and more snot and a sore throat. Am I missing something?	Infected

Table 3 Example user state sequences from real-world tweets (translated to English by a native Spanish speaker). We used HFSTM-A to classify tweets to different states. As we can see from the table, our model can capture the difference between different states and also the state transitions

4.5 Fitting flu trend using additional state information (Q. 4)

Additionally, to test the predictive capability of our models, we design a flu-case count prediction task on our test datasets, after training on *TrainData*. We compare four models: (A) the baseline model, which uses classical linear regression techniques and word counts to predict case count numbers; our models (B) HFSTM and (C) HFSTM-A, where we improve the word counts by attaching each word with the state estimated by MLE; and (D) GFT (Google Flu Trend). In all four cases we use a LASSO based linear regression model to predict the number of cases of influenza like illnesses recorded by PAHO (the ground-truth). We predict per-weekly values as both PAHO and GFT give counts only on a weekly basis.

The baseline model uses a set of features created from the counts of 114 flu related words. For *TestPeriod-1*, we count these words over $1.8M$ tweets from $0.72M$ users that were filtered by containing at least one keyword from our vocabulary (similarly for *TestPeriod-2*, *TestPeriod-3*). These word counts were then collated into a single feature vector defined as the number of tweets

containing a single word per week. We then regressed this set of counts to the PAHO case counts for each week.

Our models improve upon the baseline model by incorporating the state of the user when a word was tweeted. In this way we capture the context of a word/tweet as implied by our HFSTM and HFSTM-A models. For instance, if the word 'cold' is used in a normal conversation it probably means temperature but if it is used while a person is in the I state it is likely referring to flu related symptoms. For our models, we also use a LASSO regression to make predictions in a similar fashion. However the feature vector is created from a count of the top 20 words from each state, appended to the word of each state, such that $(cold, S)$ is counted differently from $(cold, I)$.

For GFT, we directly collect data from the Google Flu Trends website[4], and then apply the same regression as used in other methods to predict the number of infection cases. Note that as GFT is a state-of-the-art production system with highly optimized proprietary vocabulary lists, we do not expect to beat it consistently, yet as we describe later, we note some interesting results.

In all types of models the same LASSO regression is applied to the time series. For each time point a LASSO regression was fit to the last 10 weeks of data. The model was then used to predict either for zero, one, or two weeks in the future, depending on the lag; the best lag was chosen for each method. We evaluate all these methods for different countries (individually and aggregated) in South America on *TestPeriod-1*, *TestPeriod-2* and *TestPeriod-3*. We first discuss results on *TestPeriod-1* and *TestPeriod-2*, which are in the same year (2013) as the *TrainData*, then we show the qualitatively similar results on *TestPeriod-3*, which is in a different year (2014).

Fig. 6(a)–(d) show the comparison between the four models for different scenarios in 2013. Fig. 6(a) and (d) shows the aggregated cases for all countries for *TestPeriod-1* and *TestPeriod-2*, which is the same test cases we used in Chen et al. (2014). We further expand the test cases by including two example countries: Argentina and Chile for *TestPeriod-1* in Fig. 6(b) and (c). We chose Argentina and Chile as they had the largest number of tweets in our dataset. We make several observations. Firstly, as expected from the previous results, the performance of HFSTM and HFSTM-A are close to each other in all cases, despite a large vocabulary difference. The RMSE values of HFSTM-A for the four plots are $501, 437, 108, 345$ respectively, and the values for HFSTM are $485, 453, 115, 350$ respectively. The difference for our methods was only about 12. Secondly, it is clear from the figures that both HFSTM and HFSTM-A outperform the baseline method (of keyword counting) in all cases—demonstrating that the state knowledge is important and our models are carefully learning that information correctly (as a contrast to the difference between HFSTM and HFSTM-A above, the RMSE value difference between HFSTM-A and the baseline for the 4 plots are about $[210, 112, 120, 80]$ respectively). Finally, we also see that the predictions from our methods are comparable qualitatively to the state-of-the-art GFT predictions, even though

[4] `http://www.google.org/flutrends`

our methods were just implemented as a research prototype without sophisticated optimizations. In fact, although GFT performs better than HFSTM and HFSTM-A in Figures 6(a) and (b) in the RMSE scores, for Figures 6(c) and (d), our methods perform as well, and even *outperform* GFT (with an average RMSE difference of about 24). Significantly, in Figures 6(a), (c) and (d), GFT clearly overestimates the peak which our methods do not (this is an important issue with GFT which was also documented and observed in context of another US flu season as well (Butler, 2013)).

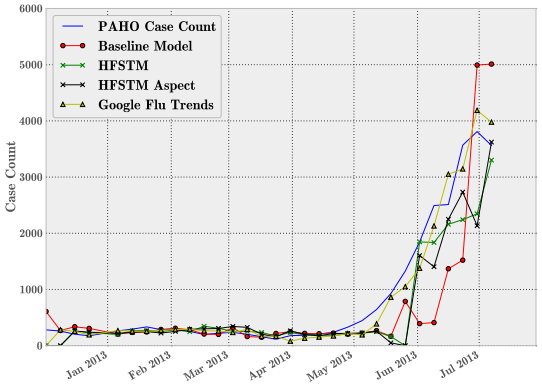
For *TestPeriod-3* in 2014, we have similar observations. The performance of HFSTM and HFSTM-A are close to each other (with 544, 599 RMSE values). GFT , although having a better RMSE value (421), clearly *overestimates* the peak. The baseline method exhibits the worst performance with an RMSE value of 871. All of the results on our test datasets show that including the epidemiological state information of users via our models can potentially benefit the prediction of infection cases dramatically.

4.6 Bridging the Social and the Epidemiological (Q.5)

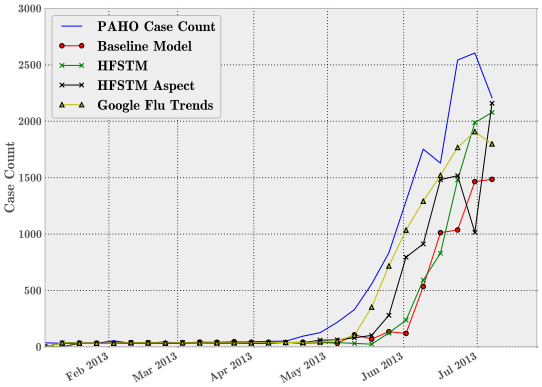
Finally, as mentioned before, another key contribution of our models is to bridge the gap between epidemiological models and social activity models. An important recent observation (Matsubara et al., 2012; Yang and Leskovec, 2011) was that the fall-part of any social activity profile is power-law—in contrast to standard epidemiological models like SEIR/SIR which give an exponential drop-off. How can they be reconciled? In the following, we show that accounting for the differences in the epidemiological state as learnt by our models, the two different activity profiles look the same i.e. they drop-off exponentially as expected from standard epidemiological models.

To test our hypothesis, we chose commonly occuring flu-keywords—such as enfermo (sick), mal (bad), fiebre (fever), dolor (pain)—for the analysis. Firstly, we count the total occurences of these keywords in *TestPeriod-1*. For each keyword we identify the falling part of its activity-curve. We then fit each curve with power law and exponential function. As expected from Matsubara et al. (2012), Fig. 8 results from HFSTM and HFSTM-A (a) and (c) both show that the power-law function provides a much better fit of the falling part of the curve compare to the exponential function (RMSE scores of exponential functions is 1.5 times higher than that of power law in both HFSTM and HFSTM-A cases).

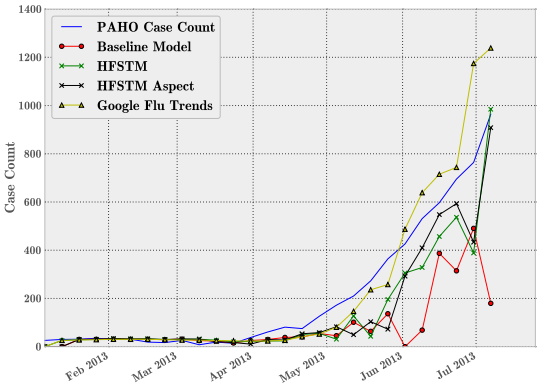
Secondly, to study the effect of our model on the activity profiles of these keywords: we count total occurrences of these keywords in the tweets which are tweeted *only by infected* users (i.e. by those users we learn as being in I). Again, we fit the falling part of each curve with a power law and a exponential function. In contrast to the previous figure, we see that now exponential fit is much better than a power law fit, the RMSE score of power law is ~ 1.9 times higher than that of exponential functions in both HFSTM and HFSTM-A cases (see Fig. 8(b) and (d))—matching what we would expect from an



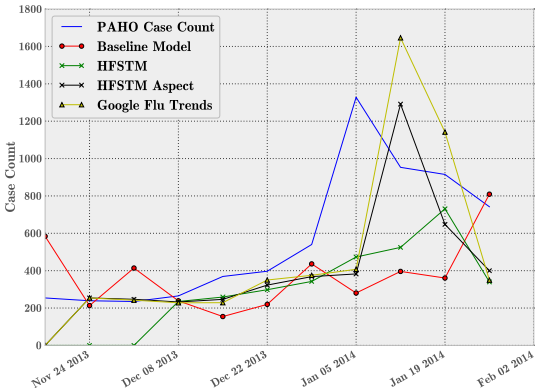
(a) All countries, *TestPeriod-1*



(b) Argentina, *TestPeriod-1*



(c) Chile, *TestPeriod-1*



(d) All countries, *TestPeriod-2*

Fig. 6 Evaluation for the two test datatsets in 2013. Comparison of the week to week predictions against PAHO case counts using the four models: baseline model, HFSTM, HFSTM-A, and GFT (Google Flu Trend). Our models outperform the baseline, performance of HFSTM and HFSTM-A are similar, and are comparable to GFT. GFT overestimates the peak in (a), (c) and (d). (a) All countries, for *TestPeriod-1*; (b) Argentina, for *TestPeriod-1*; (c) Chile, for *TestPeriod-1*; and (d) All countries, for *TestPeriod-2*

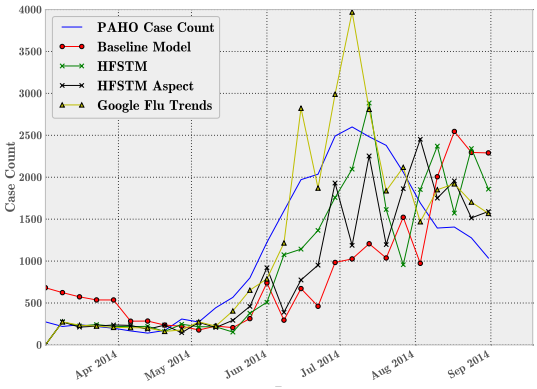


Fig. 7 Evaluation for test dataset in 2014 (*TestPeriod-3*). Comparison of the week to week predictions against PAHO case counts using the four models. The comparison is based on all countries in the dataset. We observe that the performance of HFSTM and HFSTM-A are similar and comparable to GFT, and GFT overestimates the peak.

epidemiological model like SEIS. Thus this demonstrates that finer-grained modeling can explain differences between the biological activity and the social activity which is used as its proxy.

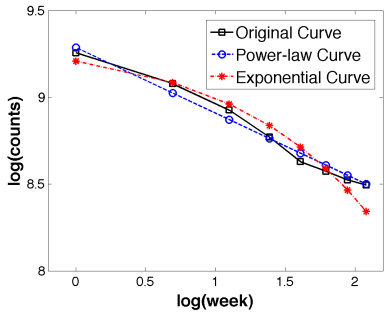
4.7 Summary of observations

In sum, the main observations from our experiments are:

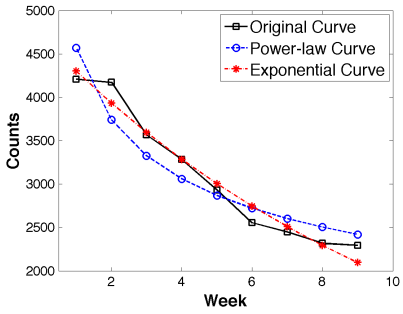
1. Our models HFSTM and HFSTM-A learn both state topic distributions and transitions, which match epidemiological intuition. The performance of HFSTM-A is robust despite of an enlarged and noisy vocabulary.
2. Our models consistently get better flu *case-count* predictions than naive vocabulary assessment (the baseline model), over datasets covering multiple time-periods.
3. Our models make better flu-*peak* predictions than Google Flu Trends for the aggregated curve in both our datasets (including individual countries like Chile).
4. Our models make qualitatively comparable flu *case-count* predictions to Google Flu Trends (even beating them in some cases).
5. Our models can potentially bridge the gap between models of biological activities and their social proxies.

5 Discussion and Conclusion

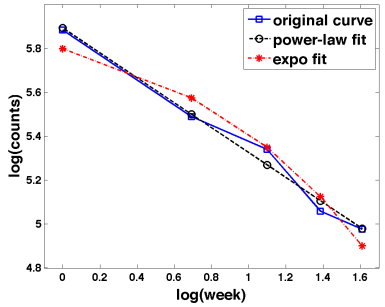
Predicting the hidden state of a user from a sequence of tweets is highly challenging. Naive methods to find the states in such sparse and large data are



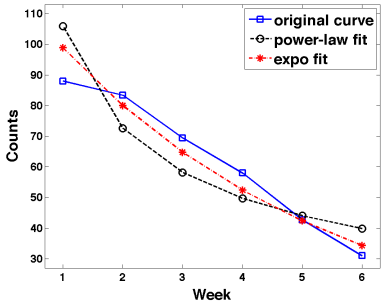
(a) Total Keyword activity (log-log) by HFSTM



(b) Keyword activity in learnt I state (lin-lin) by HFSTM



(c) Total Keyword activity (log-log) by HFSTM-A



(d) Keyword activity in learnt I state (lin-lin) by HFSTM-A

Fig. 8 Finer grained models help bridge the gap between social and epidemiological activity models. (a), (c) Power law describes keyword activity better (in *log-log* axes to show the difference); while (b), (d) Exponential function explains well the falling part of the curves for keyword activity (note the *linear* axes). The results from HFSTM and HFSTM-A agree with each other

computationally intractable. However, our proposed methods, HFSTM and HFSTM-A have the capability to use this sparseness efficiently to produce a generative model. It satisfies the requirements of low dimensional representation of the data while retaining enough information about the system. Through extensive experiments on real tweet datasets, we showed how our methods can effectively and robustly model hidden states of a user and the associated transitions, and use it to improve flu-trend prediction, including avoiding recent errors discovered in methods like Google Flu Trends. Further, our models use public data, and our results were stable across two *different* time-periods. We also showed how our model can reconcile seemingly different behaviors from social and epidemiological models.

As mentioned in the introduction, current approaches for predicting flu using information gleaned from the Twitter data are often devoid of any epidemiological significance and hence there is a great chasm between the data driven flu trend modelling using Twitter data and the model-based, simulation-oriented epidemiological models such as SI, SIR and SEIS. Hence more broadly,

our technique can act as the missing link between this apparently uncorrelated line of research—lending a state aware nature to data-driven models and simultaneously, can let simulation oriented models estimate their state transition matrices by maximizing data likelihood.

6 Future Work

We have several directions to further extend our work.

The state transitions probabilities our models learn can be used to estimate parameters in traditional epidemiological models, such as the transmission rate, the removal rate, the infection prevalence threshold, etc. We can study how these epidemiological parameters behave in the context of twitter, and how these social-media-derived parameters reflect on the real situation.

Secondly, as twitter is a highly connected social network, we can integrate the network structure into our models and improve the results. Currently our models assume independency between twitter users, and estimate a user's states by only looking at his/her own tweets. However in reality, people are more likely to get infected if most of their friends are infected. Hence the neighbors of a node in the network have some bias on the state transition of the node, which can be integrated into our models.

Thirdly, we can improve the efficiency of the inference algorithms. Note that our algorithms are practical enough to run on real world datasets as used in this paper. Nevertheless it will be interesting to improve the scalability of our approach by exploring approaches like distributed EM, or other inference algorithms like MCMC.

Finally, our proposed methods are general enough and can be easily extended to other domains such as monitoring organized protests where a user can go through several states of protest like 'not interested', 'ambivalent', 'active', 'resigned', etc.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1353346, by the Maryland Procurement Office under contract H98230-14-C-0127, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, and by the VT College of Engineering. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the respective funding agencies.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707.

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans*. Oxford University Press.

Andrews, M. and Vigliocco, G. (2010). The hidden markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, 2(1):101–113.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576.

Beretta, E. and Takeuchi, Y. (1995). Global stability of an SIR epidemic model with time delays. *Journal of Mathematical Biology*, 33(3):250–260.

Blasiak, S. and Rangwala, H. (2011). A Hidden Markov Model Variant for Sequence Classification. In *the 21nd International Joint Conference on Artificial Intelligence*, pages 1192–1197.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic Topic Models. *Signal Processing Magazine, IEEE*, 27(6):55–65.

Blei, D. and Lafferty, J. (2006). Dynamic Topic Models. In *the 23rd International Conference on Machine Learning*, pages 113–120.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brennan, S. P., Sadilek, A., and Kautz, H. A. (2013). Towards understanding global spread of disease from everyday interpersonal interactions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2783–2789. AAAI Press.

Butler, D. (2013). When Google got Flu Wrong. *Nature*, 494(7436):155–156.

Chakraborty, P., Khadivi, P., Lewis, B., Mahendiran, A., Chen, J., Butler, P., Nsoesie, E., Mekaru, S., Brownstein, J., Marathe, M., and Ramakrishnan, N. (2014). Forecasting a moving target: Ensemble models for ili case count predictions. In *2014 SIAM International Conference on Data Mining*, SDM '14.

Chen, L., Hossain, K. S. M. T., Butler, P., Ramakrishnan, N., and Prakash, B. A. (2014). Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '14.

Christakis, N. A. and Fowler, J. H. (2010). Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*, 5(9):e12948.

Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*,

pages 115–122. ACM.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2008). Detecting Influenza Epidemics using Search Engine Query Data. *Nature*, 457(7232):1012–1014.

Glance, N., Hurst, M., and Tomokiyo, T. (2004). Blogpulse: Automated trend discovery for weblogs. *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*, 2004.

Gruber, A., Weiss, Y., and Rosen-Zvi, M. (2007). Hidden topic markov models. In *International Conference on Artificial Intelligence and Statistics*, pages 163–170.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *Society for Industrial and Applied Mathematics, SIAM review*, 42(4):599–653.

Hong, L., Yin, D., Guo, J., and Davison, B. (2011). Tracking Trends: Incorporating Term Volume into Temporal Topic Models. In *the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 484–492.

Jacquez, J. and Simon, C. (1993). The stochastic SI model with recruitment and deaths I. Comparison with the closed SIS model. *Mathematical biosciences*, 117(1):77–125.

Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 789–795.

Lampos, V. and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72.

Lampos, V., De Bie, T., and Cristianini, N. (2010). Flu detector: Tracking epidemics on twitter. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 599–602.

Lazer, D. M., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205.

Lee, K., Agrawal, A., and Choudhary, A. (2013). Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 1474–1477. ACM.

Li, J. and Cardie, C. (2013). Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.

Li, M. and Muldowney, J. (1995). Global stability for the seir model in epidemiology. *Mathematical Biosciences*, 125(2):155–164.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2012). Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 6–14.

PAHO (2012). Epidemic disease database, pan american health organization. `http://ais.paho.org/phip/viz/ed_flu.asp`.

Paul, M. and Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 265–272.

Paul, M. and Girju, R. (2010). A Two-dimensional Topic-aspect Model for Discovering Multi-faceted Topics. *Urbana*, 51:61801.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA. ACM.

Sadilek, A., Kautz, H., and Silenzio, V. (2012). Predicting disease transmission from geo-tagged micro-blog data. In *AAAI Conference on Artificial Intelligence*.

Spasojevic, N., Yan, J., Rao, A., and Bhattacharyya, P. (2014). Lasta: Large scale topic assignment on multiple social networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1809–1818, New York, NY, USA. ACM.

Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. (2004). Probabilistic Author-topic Models for Information Discovery. In *The 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 306–315.

Wang, X. and McCallum, A. (2006). Topics Over Time: a non-Markov Continuous-time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on web search and data mining*, pages 177–186. ACM.

Yang, J., McAuley, J., Leskovec, J., LePendu, P., and Shah, N. (2014a). Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 783–794.

Yang, S.-H., Kolcz, A., Schlaikjer, A., and Gupta, P. (2014b). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1907–1916, New York, NY, USA. ACM.

Zhao, S., Zhong, L., Wickramasuriya, J., and Vasudevan, V. (2011). Human as real-time sensors of social and physical events: A case study of twitter and sports games. *arXiv preprint arXiv:1106.4300*.

Appendix A HFSTM-A-FIT

In this appendix, we show the equations we designed for HFSTM-A-FIT. Note that the outlines of the HFSTM-FIT algorithm is similar to HFSTM-A-FIT, one can derive equations for HFSTM-FIT from the content we show below.

Let K, T, N, and U be the number of states, number of tweets per user, number of words per tweet, and total number of users. Let $O =< O_1, O_2, \ldots, O_T >$ and $S =< S_1, S_2, \ldots, S_T >$ the observed sequences of tweets and hidden states respectively for a particular user.

Here is a list of symbols that we will use.

1. ϵ: the prior for the binary state switching variable, which determines whether state of a tweet is drawn from the transition probability matrix or simply copied from the state of the previous tweet (a number in $(0, 1]$)
2. π: initial state probability (size is $1 \times K$)
3. η: tansition probability matrix (size is $K \times K$)
4. ϕ: word distrtibution for each state (size is $K \times W$, where W is the total number of keywords for all of the states)
5. w_{tn}: the nth word in the tth tweet
6. λ: the background switch variable
7. c: the topic switch variable
8. y: the observed aspect value

For HFSTM-A, as mentioned in Section 3.3, the value of λ is biased by the observed aspect value y. We use λ instead of λ_y in the following for brevity, but remember the λ value in the equations is actually calculated using:

$$\begin{aligned}
\lambda_{y_i=0} &= \lambda \\
\lambda_{y_i=1} &= \lambda + b * (1 - \lambda) \\
c_{y_i=0} &= c - a * c \\
c_{y_i=1} &= c + a * (1 - c)
\end{aligned}$$

We want to learn all the parameters given the tweet sequence. For compact notation we use $H = (\epsilon, \pi, \eta, \phi, \lambda, c)$. In HFSTM-A-FIT, we use forward backward procedure for which we define forward variable $A_t(i)$ and backward variable $B_t(i)$ as follows.

$$A_t(i) = P(O_1, O_2, \ldots, O_t, S_t = i|H)$$

$$B_t(i) = P(O_{t+1}, \ldots, O_T|S_t = i, H)$$

Let $\gamma_t(i)$ be the probability of being in state S_i at for tth tweet given the observed tweet sequence O and other model parameters. For each user the size of γ is $2K \times T$ (with the first K states as the states which are copies of the previous state, and the second K states which are derived after a transition). This probability can be expressed by the forward and backward probabilities.

$$\begin{aligned}
\gamma_t(i) &= P(S_t = i|O, H) \\
&= \frac{A_t(i)B_t(i)}{P(O|H)} \\
&= \frac{A_t(i)B_t(i)}{\sum_{i=1}^{2K} A_t(i)B_t(i)}
\end{aligned}$$

We have two switch variables in the model: l, x. If $l = 1$, the word is generated either by states or topics, if $l = 0$ it's generated by background. If $x = 0$, the word is generated by topics, if $x = 1$ it's by states.

For $l_i = 1$, which means that w_i is generated by either state or topics.

$$P(l_i = 1|\lambda, c, H, w) = \frac{P(l_i = 1|\lambda, c, H)P(w|l_i = 1, \lambda, c, H)}{P(w|\lambda, c, H)}$$
$$= \frac{\lambda P(w_i|\lambda, c, H, l_i = 1, w_{-i})P(w_{-i}|\lambda, c, H, l_i = 1)}{P(w_i|\lambda, c, H, w_{-i})P(w_{-i}|\lambda, c, H)}$$
$$= \frac{\lambda \sum_{x_i}[P(w_i|\lambda, c, H, l_i = 1, x_i, w_{-i})P(x_i|\lambda, c, H, l_i = 1, w_{-i})]}{\sum_{l_i}[P(w_i|\lambda, c, H, l_i, w_{-i})P(l_i|\lambda, c, H, w_{-i})]}$$
$$= \frac{\lambda[(\sum_{topic}\phi_{topic}(w_i)P(topic|x_i=0,l_i=1,\lambda,c,H,w_{-i}))(1-c)+(\sum_{state}\phi_{state}(w_i)\gamma_i(state))c]}{\lambda[(\sum_{topic}\phi_{topic}(w_i)P(topic|\ldots))(1-c)+(\sum_{state}\phi_{state}(w_i)\gamma_i(state))c]+(1-\lambda)\phi_{Bak}(w_i)}$$

For $l_i = 0$, w_i is generated by background.

$$P(l_i = 0|\lambda, c, H, w) = \frac{P(l_i = 0|\lambda, c, H)P(w|l_i = 0, \lambda, c, H)}{P(w|\lambda, c, H)}$$
$$= \frac{(1-\lambda)\phi_{Bak}(w_i)}{\lambda[(\sum_{topic}\phi_{topic}(w_i)P(topic|\ldots))(1-c)+(\sum_{state}\phi_{state}(w_i)\gamma_i(state))c]+(1-\lambda)\phi_{Bak}(w_i)}$$

For $x_i = 0$, w_i is generated by topics.

$$P(x_i = 0|\lambda, c, H, w) = \frac{P(x_i = 0|\lambda, c, H)P(w|x_i = 0, \lambda, c, H)}{P(w|\lambda, c, H)}$$
$$= \frac{(1-c)P(w_i|\lambda, c, H, x_i = 0, w_{-i})P(w_{-i}|\lambda, c, H, x_i = 0)}{P(w_i|\lambda, c, H, w_{-i})P(w_{-i}|\lambda, c, H)}$$
$$= \frac{(1-c)\sum_{l_i}[P(w_i|\lambda, c, H, x_i = 0, l_i, w_{-i})P(l_i|\lambda, c, H, x_i = 0, w_{-i})]}{\sum_{x_i}P(w_i|\lambda, c, H, w_{-i}, x_i)P(x_i|\lambda, c, H, w_{-i})}$$
$$= \frac{(1-c)[(\sum_{topic}\phi_{topic}(w_i)P(topic|x_i=0,l_i=1,\lambda,c,H,w_{-i}))\lambda+\phi_{Bak}(w_i)(1-\lambda)]}{(1-c)[(\sum_{top}\phi_{top}(w_i)P(top|\ldots))\lambda+\phi_{Bak}(w_i)(1-\lambda)]+c[(\sum_{sta}\phi_{sta}(w_i)\gamma_i(sta))\lambda+\phi_{Bak}(w_i)(1-\lambda)]}$$

For $x_i = 1$, w_i is generated by states.

$$P(x_i = 1|\lambda, c, H, w) = \frac{P(x_i = 1|\lambda, c, H)P(w|x_i = 1, \lambda, c, H)}{P(w|\lambda, c, H)}$$
$$= \frac{c[(\sum_{sta}\phi_{sta}(w_i)\gamma_i(sta))\lambda+\phi_{Bak}(w_i)(1-\lambda)]}{(1-c)[(\sum_{top}\phi_{top}(w_i)P(top|\ldots))\lambda+\phi_{Bak}(w_i)(1-\lambda)]+c[(\sum_{sta}\phi_{sta}(w_i)\gamma_i(sta))\lambda+\phi_{Bak}(w_i)(1-\lambda)]}$$

Forward variable: We now further expand the forward variable in more details. The Initialization is as follows:

For $1 \leq i \leq K$:

$$\begin{aligned}
A_1(i) &= P(O_1, S_1 = i|H) \\
&= P(O_1|S_1 = i, H)P((S_1 = i|H) \\
&= \pi_i \prod_{n=1}^{N} P(w_{1n}|S_1 = i, H) \\
&= \pi_i \prod_{n=1}^{N} \{(1-\lambda)\phi_{Bak}(w_{1n}) + \lambda[(1-c)\sum_{top}\phi_{top}(w_{1n})P(top|\ldots) + c\phi^i(w_{1n})]\}
\end{aligned}$$

For $K+1 \leq i \leq 2K$: $A_1(i) = 0$

Induction is as follows:

For $1 \leq j \leq K$:

$$\begin{aligned}
A_t(j) &= P(O_1, O_2, \ldots, O_t, S_t = j|H) \\
&= (\sum_{i}^{2K} A_{t-1}(i)\epsilon\eta_{ij})P(O_t|S_t = j, H) \\
&= (\sum_{i}^{2K} A_{t-1}(i)\epsilon\eta_{ij}) \prod_{n=1}^{N} \{(1-\lambda)\phi_{Bak}(w_{1n}) \\
&+ \lambda[(1-c)\sum_{top} \phi_{top}(w_{1n})P(top|\ldots) + c\phi^j(w_{1n})]\}
\end{aligned}$$

For $K+1 \leq j \leq 2K$:

$$\begin{aligned}
&A_t(j) = P(O_1, O_2, \ldots, O_t, S_t = j|H) \\
&= (A_{t-1}(j) + A_{t-1}(j-K))(1-\epsilon) \prod_{n=1}^{N} \{(1-\lambda)\phi_{Bak}(w_{1n}) \\
&+ \lambda[(1-c)\sum_{top} \phi_{top}(w_{1n})P(top|\ldots) + c\phi^j(w_{1n})]\}
\end{aligned}$$

Backward variable: The initialization for backward variable is as follows:
For $1 \leq i \leq 2K$:

$$B_T(i) = 1$$

Induction is as follows:
For $1 \leq i \leq K$:

$$\begin{aligned}
&B_t(i) = P(O_{t+1}, \ldots, O_T|S_t = i, H) \\
&= \left(\sum_{j}^{K} \epsilon\eta_{ij} P(O_{t+1}|S_{t+1} = j, H)B_{t+1}(j)\right) \\
&\quad + (1-\epsilon)P(O_{t+1}|S_{t+1} = i+K, H)B_{t+1}(i+K) \\
&= \left(\sum_{j}^{K} \epsilon\eta_{ij} \prod_{n=1}^{N} \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n})P(top|\ldots)\right. \\
&\quad \left. + c\phi^j(w_{(t+1)n})]\}B_{t+1}(j)\right) + (1-\epsilon) \prod_{n=1}^{N} \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) \\
&\quad + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n})P(top|\ldots) + c\phi^i(w_{(t+1)n})]\}B_{t+1}(i+K)
\end{aligned}$$

For $K+1 \leq i \leq 2K$:

$$
\begin{aligned}
&B_t(i) = P(O_{t+1}, \ldots, O_T | S_t = i, H) \\
&= \left(\sum_j^K \epsilon \eta_{ij} P(O_{t+1} | S_{t+1} = j, H) B_{t+1}(j) \right) \\
&\quad + (1-\epsilon) P(O_{t+1} | S_{t+1} = i, H) B_{t+1}(i) \\
&= \Bigg(\sum_j^K \epsilon \eta_{ij} \prod_{n=1}^N \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n}) P(top|\ldots) \\
&\quad + c\phi^j(w_{(t+1)n})]\} B_{t+1}(j) \Bigg) + (1-\epsilon) \prod_{n=1}^N \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) \\
&\quad + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n}) P(top|\ldots) + c\phi^{i-K}(w_{(t+1)n})]\} B_{t+1}(i)
\end{aligned}
$$

Define z as follows:

$$
\begin{aligned}
z_{t,n}(i) &= P(T_{tn} = i | l_{tn} = 1, x_{tn} = 0, w_{tn}, H) \\
&= \frac{P(w_{tn} | T_{tn} = i, H, l_{tn} = 1, x_{tn} = 0) P(T_{tn} = i | l_{tn} = 1, x_{tn} = 0, H)}{P(w_{tn} | l_{tn} = 1, x_{tn} = 0, H)} \\
&= \frac{\phi_{top=i}(w_{tn}) P(T_{tn} = i | l_{tn} = 1, x_{tn} = 0, H)}{\sum_i [\phi_{top=i}(w_{tn}) P(T_{tn} = i | l_{tn} = 1, x_{tn} = 0, H)]}
\end{aligned}
$$

Let $\xi_t(i,j)$ be the probability of being in state S_i at time t, and state S_j at time $t+1$, given O and other model parameters.

$$
\begin{aligned}
\xi_t(i,j) &= P(S_t = i, S_{t+1} = j | O, H) \\
&= \frac{P(S_t = i, S_{t+1} = j, O | H)}{P(O|H)}
\end{aligned}
$$

To express $\xi_t(i,j)$, we have the following definition.
For $1 \le i \le 2K$ and $1 \le j \le K$:

$$
\begin{aligned}
T_1 &= A_t(i) \epsilon \eta_{ij} P(O_{t+1} | S_{t+1} = j, H) B_{t+1}(j) \\
&= A_t(i) \epsilon \eta_{ij} \prod_{n=1}^N \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) \\
&\quad + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n}) P(top|\ldots) + c\phi^j(w_{(t+1)n})]\} B_{t+1}(j)
\end{aligned}
$$

For $1 \le i \le K$ and $K+1 \le j \le 2K$:

$$
\begin{aligned}
T_2 &= A_t(i)(1-\epsilon) P(O_{t+1} | S_{t+1} = i+K, H) B_{t+1}(i+K) \\
&= A_t(i)(1-\epsilon) \prod_{n=1}^N \{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) \\
&\quad + \lambda[(1-c)\sum_{top} \phi_{top}(w_{(t+1)n}) P(top|\ldots) + c\phi^i(w_{(t+1)n})]\} B_{t+1}(i+K)
\end{aligned}
$$

For $K+1 \leq i \leq 2K$ and $K+1 \leq j \leq 2K$:

$$\begin{aligned} T_3 &= A_t(i)(1-\epsilon)P(O_{t+1}|S_{t+1}=i,H)B_{t+1}(i) \\ &= A_t(i)(1-\epsilon)\prod_{n=1}^{N}\{(1-\lambda)\phi_{Bak}(w_{(t+1)n}) \\ &\quad + \lambda[(1-c)\sum_{top}\phi_{top}(w_{(t+1)n})P(top|\ldots) + c\phi^{i-K}(w_{(t+1)n})]\}B_{t+1}(i) \end{aligned}$$

Correspondingly, we have the following ξ values according to the different i, j value range:

$$\xi_t(i,j) = \frac{T_1}{\sum_i \sum_j (T_1+T_2+T_3)}$$
$$\xi_t(i,j) = \frac{T_2}{\sum_i \sum_j (T_1+T_2+T_3)}$$
$$\xi_t(i,j) = \frac{T_3}{\sum_i \sum_j (T_1+T_2+T_3)}$$

Estimation of parameters:
We use the following equations to estimate the parameter values in the M-step.
For estimating ϵ:

$$\epsilon = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{i=1}^{2K}\sum_{j=1}^{K}\xi(i,j)}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{i=1}^{2K}\sum_{j=1}^{2K}\xi(i,j)}$$

For estimating π:

$$\pi_i = \frac{\sum_{u=1}^{U}\gamma_1(i)}{\sum_{u=1}^{U}\sum_{i=1}^{K}\gamma_1(i)} \quad \text{for } 1 \leq i \leq K$$

For estimating η:

$$\eta_{ij} = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}(\xi_t(i,j)+\xi_t(i+K,j))}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{j=1}^{K}(\xi_t(i,j)+\xi_t(i+K,j))} \quad \text{for } 1 \leq i \leq K,\ 1 \leq j \leq K$$

For estimating λ:

$$\lambda = \frac{\sum_u \sum_t \frac{1}{N_t}\sum_{n=1}^{N_t} P(l_{tn}=1|\lambda,c,H,w)}{UT}$$

For estimating c:

$$c = \frac{\sum_u \sum_t \frac{1}{N_t}\sum_{n=1}^{N_t} P(l_{tn}=1|\lambda,c,H,w)P(x_{tn}=1|\lambda,c,H,w)}{\sum_u \sum_t \frac{1}{N_t}\sum_{n=1}^{N_t} P(l_{tn}=1|\lambda,c,H,w)}$$

For estimating ϕ:

$$\phi_(^i w) = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{\substack{1\leq n\leq N \\ \& \\ w=w_{tn}}} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=1|\lambda,c,H,O)(\gamma_t(i)+\gamma_t(i+K))}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{w=1}^{W}\sum_{\substack{1\leq n\leq N \\ \& \\ w=w_{tn}}} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=1|\lambda,c,H,O)(\gamma_t(i)+\gamma_t(i+K))}$$

for $1 \leq i \leq K$

$$\phi_{Bak}(w) = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{\substack{1\le n\le N \\ \& \\ w=w_{tn}}} P(l_{tn}=0|\lambda,c,H,O)}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{w=1}^{W}\sum_{\substack{1\le n\le N \\ \& \\ w=w_{tn}}} P(l_{tn}=0|\lambda,c,H,O)}$$

$$\phi_{Topic}(w) = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{\substack{1\le n\le N \\ \& \\ w=w_{tn}}} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=0|\lambda,c,H,O)z_{t,n}(Topic)}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{w=1}^{W}\sum_{\substack{1\le n\le N \\ \& \\ w=w_{tn}}} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=0|\lambda,c,H,O)z_{t,n}(Topic)}$$

$$P(T_{tn}=i|l_{tn}=1,x_{tn}=0,H) = \frac{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{n=1}^{N_t} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=0|\lambda,c,H,O)z_{t,n}(i)}{\sum_{u=1}^{U}\sum_{t=1}^{T}\sum_{n=1}^{N_t} P(l_{tn}=1|\lambda,c,H,O)P(x_{tn}=0|\lambda,c,H,O)}$$