Auto-Generated User Profile Schemas as a Lens for Understanding Product Experience Trajectories on Reddit

Xinru Yan, Yohan Jo, Carolyn Rose

CMU-LTI-19-006

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Auto-Generated User Profile Schemas as a Lens for Understanding Product Experience Trajectories on Reddit

Xinru Yan, Yohan Jo, Carolyn Rose
Language Technologies Institute
Carnegie Mellon University
xinruyan, yohanj, cprose@cs.cmu.edu

Abstract

Product reviews offer a spectrum of snap-shot perspectives on product experiences. In contrast, product oriented discussion forums, such as found on Reddit, offer the possibility to examine reactions to products in the context of a user's posting history. Examination of typical product usage trajectories affords extraction of more nuanced understandings of user experiences that connect opinions on products with features of users or situations in which products are used. This paper presents a computational modeling approach to construction of user profiles that enable construction of structured representations of user experiences with products over time. Using this approach it is possible to identify types of experience trajectories and how they are differentially associated with characteristics of users.

Introduction

The contribution of this paper is a novel computational approach to construction of user profiles along with a method for using them as a lens for understanding experience trajectories over time for types of users. We illustrate the use of this lens in connection with cosmetic products as they are discussed in Reddit[1] discussion forums.

One of the oldest and best established forms of text mining is opinion mining (Pang, Lee, and others 2008; Liu 2012), which has attracted much attention both as a research topic and as an industry-relevant technology with tremendous market potential. The bulk of foundational work in this area has taken a decontextualized approach to extraction of user opinions about products, most typically by extracting opinions from product reviews, which give little contextual information about the characteristics of the people writing the reviews or what their past experiences have been. Latent user types have played a role in modeling user preferences through the common collaborative filtering and matrix factorization approaches used frequently in recommender systems (Schafer et al. 2007; Ekstrand et al. 2011). A key idea here is that there are kinds of users, and kinds of users respond similarly across products. What this type of approach does not offer is a time-series view of how product usage is contextualized in trajectories of user experiences over time.

In contrast, typical social media platforms such as Twitter[2], Facebook[3] , Tumblr[4] and Reddit afford users the opportunity to project a constructed identity by means of their reported experiences and views as their lives unfold over time. This rich form of user history affords the opportunity to make sense of their expressed opinions in a contextually informed way. Users disclose many details about their lived experiences through posts, which may include text, links, images and videos. Some social media platforms such as Facebook also provide structured profile forms for users to fill out with information such as gender, age, and education (Mislove et al. 2010; Li, Ritter, and Hovy 2014; Wang et al. 2018). An abstraction over user profiles used as a lens for modeling post trajectories over time would no doubt be of tremendous value. However, frequently users neglect to fill out profile information, and many users do not post frequently or continue to post over extensive periods of time. Thus, in our work, we have developed an approach that is robust to this variation, allowing story schemas to be induced using a method that requires only the most frequently offered user data (i.e., posts), and which can be constructed over the rich data provided by extensive posters.

A **story schema** is a data structure with slots that represent typical elements of a story. It can be used for extraction of important details from text that has been aggregated at different time scales, such as a week, a year, or a user's whole posting history. In our work we first induce the concept of a story schema, which is induced from within-post structure and then applied to whole post histories to construct a representation of a user, which can be thought of as the "story" of that user's life. We will refer to this as a constructed **user profile**. We then segment user post histories into time periods and again apply the schema in order to construct a representation of types of time periods, or **user states**. Finally, we utilize the constructed user states to build state transition diagrams that represent typical **user trajectories** for each type of user.

The remainder of the paper is organized as follows: First,

[1] https://www.reddit.com/

[2] https://twitter.com/

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

[4] https://www.tumblr.com/

we discuss related work in opinion mining and extracting user profiles on social media, background on similar modeling approaches and motivate our approach, which leverages past work on auto-constructed discussion schemas. Next we describe our data set, task, and computational pipeline. Following that, we illustrate how our method can be applied in order to understand user trajectories as they are portrayed in Reddit discussion forums focused on cosmetic usage. We conclude with a discussion of the limitations of the approach and plans for continued work.

Background & Related Work

In this section we introduce the prior work on opinion mining in social media and user profile extraction. We also discuss research on Reddit as the platform where we extract our data, and prior computational work that offers tools to aid in addressing our computational challenges.

Opinion Mining and User Profile Extraction on Social Media

With the dramatic growth of social media platforms such as blog and microblog networks, review sites and discussion forums, opinion mining as a powerful tool for analyzing social media data has gained incredible attention. Researchers have approached this problem as a binary classification task, i.e., detecting positive or negative sentiment on user generated text. For example, Liang and Dai proposed a system to automatically extract opinions from tweets and analyze their sentiment. Mei et al. developed an HMM based model to extract the mixture of topics and sentiment expressions simultaneously on Weblog data. Penalver-Martinez et al. utilized a feature-based and vector-based method to evaluate sentiment in online movie reviews. O'Connor et al. analyzed political opinions on contemporaneous Twitter messages. It is worth mentioning that all of these efforts have taken a decontextualized approach to extraction of user opinions. In contrast, we focus on extraction of users' entire post histories to ensure that we have the contextual information about users' characteristics and their experiences.

Techniques for extraction of user profiles have evolved since the 1990s, beginning with work on recommender systems. More recently, with the increasing popularity of social media platforms, more studies focus on extraction of user profiles from the massive amount of information available in unstructured user traces. Multiple studies have treated user profile inference as a classification problem where trained models predict user characteristics, such as demographic variables, from a different types of trace data. For example, Mislove et al. collected data from two different user networks on Facebook and then exploited explicit user-provided profile data in order to predict missing profile data for other within-network users based on the network configuration. In particular, college major and year of matriculation are examples of inferred profile information. Rao and Yarowsky took a sociolinguistic approach to feature space design for Support Vector Machine (SVM) models to learn to automatically identify user attributes including gender, age, region of origin and political orientation from Twitter data. From another angle, research on user profile extraction can differ in terms of which and how many data sources are involved. Most work has focused on single-source data including the aforementioned ones. In contrast, Farseev and Chua proposed their first studies on individual wellness profiling. They infer wellness related attributes such as the BMI and trend of BMI for a user by integrating sensor data with what can be extracted from various social platforms. Wang et al. addressed the cross-media user profile extraction problem by learning user embeddings from two networks, the user-word network and user-user network. Then they used these embeddings to train models for gender classification and age regression tasks. Li, Ritter, and Hovy presented a weakly supervised framework utilizing both text features and network features for user attribute (job, spouse and education) inference on Twitter. While predicting user attributes based on their tweets, they also integrated information extracted from the users' linked Google Plus and Facebook accounts.

This prior work relies heavily on extracting factual user attributes from explicit profiles filled out by users, rather than using raw text streams (e.g., their posts). Even when raw text streams have been involved, they have been used mainly for extraction of explicitly reported user characteristics, such as birthdays, age, gender, and marital status. Wherever explicit profile forms have been made available to users, only a minority of users fill them in. Raw discussion data is more plentiful. In addition, the insight about users that can be extracted from that raw data goes beyond what is encompassed within typical user profile forms, as will be exemplified in our work reported in this paper. In particular, instead of focusing on extraction of demographic variables, we propose to induce story schemas from user posts to form user profiles that characterize users in terms of the kinds of experiences they have had in order to illuminate the attitudes they have expressed towards products in their discussions over time.

The Reddit Platform

Leading social media platforms such as Reddit, Tumblr, Facebook and Twitter have gained tremendous popularity in the recent decades, each affording a distinctive type of interaction and engagement. Reddit is a community driven website where users mainly communicate via creating and commenting on posts. Users can form their own communities, called *subreddits*, with a specific topic focus. In our work, we draw attention to two subreddits that are related to the cosmetic domain, */r/MakeupAddiction* and */r/SkincareAddiction*. Reddit contains a massive volume of discussions on a myriad of topics where users share their experiences and knowledge. This tremendous breadth makes it suitable for our purposes. Beyond the scope of this paper, while we focus on cosmetic subforums on Reddit, the goal is to develop a pipeline that could be applied to data extracted from Reddit for any selected topic within its scope.

A variety of studies in computational social science have already involved data extracted from Reddit. For example, De Choudhury and De focused on mental illness communities on Reddit. By building language models and statistical

models on self-disclosed data drawn from Reddit, they analyzed characteristics revealed in mental health social support and explored the role of social media in behavioral therapy. Fiesler et al. reported an ecosystem of community created rules over a large number of subreddits. They found that although rules on these subreddits are context dependent, they also share some common traits across the site. Tan studied how new communities emerge from old communities on Reddit. In their work, they treated each community as an entity, identified parents of communities and built genealogy graphs for communities. Gjurković and Šnajder presented a personality prediction study on Reddit where they constructed a large scale dataset with personality labels, which was then used to train and evaluate personality predictions.

To the best of our knowledge, there is no prior work on extracting user profiles as we characterize them on Reddit. Reddit does not require users to fill out structured user profiles with information such as age, gender and etc. While this lack of explicit labels would serve as a hindrance to classification-based approaches to profile construction, it is not a problem for our work involving induction of story schemas, which are then used as a lens for understanding user engagement with products over time.

Schema Induction Models

A family of models that are useful for automatic schema extraction is topic models, especially Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Griffiths, Steyvers, and Tenenbaum 2007). Topic models are probabilistic graphical models that have been widely exploited to automatically identify themes from a set of documents, as defined by distributions of frequently co-occurring words. They are sometime thought of as capturing the major points in text. Generative models such as LDA are advantageous to use for modeling because they afford the opportunity to use insights about the data and the structure of the desired inference to be incorporated into the model through the specification of statistical assumptions.

In our work, we assume that there are typical ways in which users describe their experiences and stories on Reddit. If we can induce story schemas, we can assemble details extracted from stories into a coherent representation, which in our case is utilized to form user profiles. Building on prior work in extraction of story schemas from newspaper articles (Barzilay and Lee 2004), the aim is to identify typical sequences of story elements. The story elements are often associated with characteristic sentence structures. For example, a story might introduce a character, then tell a problem the character is having, then tell how the character tried to solve the problem, and then report the resolution, and each of these may frequently be indicated through inclusion of characteristic structural elements. These structural elements, when they occur, give some indication of what to expect next. In our case, in a discussion forum when a user authors a post, they might greet the community first, then give some background about themselves, then state the topic they want to discuss and elaborate on the topic, and then seek for opinions.

Therefore, in our work, we are not only interested in the content of the text, but more importantly the structures found within the text. Conversation models (Ritter, Cherry, and Dolan 2010; Lee et al. 2013) – a subset of topic models – are designed to automatically identify forms of sentences or structural elements within utterances, and how they are typically ordered within a conversation, thus they are suitable for our purpose to identify elements of stories that are told during conversation. Several conversation models are available (Paul 2012; Wallace et al. 2013; Jo et al. 2017). In our work, we choose a recent model that offers the best performance at conversation element labeling, namely CSM (the content word filtering and speaker preferences model). In particular, CSM aims to identify various linguistic structures in sentences in given conversations. Linguistic structures refer to typical functional forms that convey diverse content, examples of which include certain speech acts, e.g., asking questions and greeting, and domain-specific message types, e.g., an error message template in technical forums (Jo et al. 2017).

CSM is a generative model of conversation, where a conversation is a sequence of utterances by speakers. The model assumes that there are a set of sentence structures, and each sentence can take one of them. Each structure is represented as a language model, i.e., a probability distribution over words. At a higher level, there are a set of states, where each state is a probability distribution over sentence-level language models. In other words, the probability for the appearance of a sentence level structure depends upon the state. Many utterances that have the same state are likely to have sentences with the same or similar structure. In addition to all these components related to sentence structure, there are also a set of content topics, each of which is also a language model. There is assumed to be a global probability distribution over content topics, which represents the probabilistic proportions of content topics.

CSM can be thought of as a combination of HMM and topic model, but adopts a deliberate design choice different from other unsupervised models that identify content-wise topics. The model captures linguistic structures using three mechanisms. First, the model encodes that, in a conversation the content being discussed transitions more slowly than the structures that convey the content, e.g., in a series of conversation turns between two speakers, one asking questions and the other answering them, the structures switch in every turn between asking and answering, but the content being discussed may remain constant. As such, the model de-emphasizes words that occur consistently throughout a conversation and identifies various co-occurrence patterns of other fast-changing words, which are likely to constitute structures. According to the design of the model, since sentences in an utterance can have different structure language models but only the same content topic, structure language models tend to learn fast-changing words and the content topic relatively constant words.

Second, CSM encodes that the structures of sentences in an utterance are probabilistically conditioned on those of the preceding utterance via states. This assumption is to capture the tendency that the structure of an utterance influences the selection of structure for the following utterance, e.g., asking

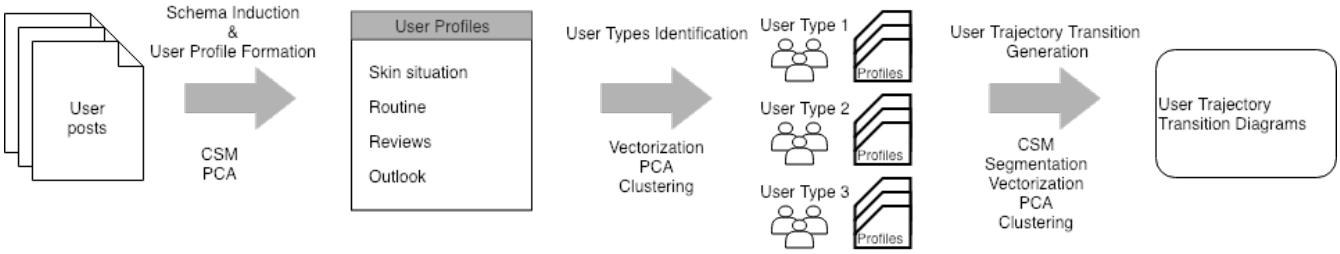

Figure 1: Computational Pipeline. Steps are on the top of the arrow and techniques are at the bottom.

a question is likely to be followed by answering the question, and greetings by greetings. As a result, the model learns linguistic structures that account for the dynamics of utterances in given conversations.

Third, the model encodes that speakers have preferences over certain structures in their utterances. For example, a speaker may ask a lot of questions in the conversation, and another speaker may mainly moderate the conversation. Modeling preferences helps the model identify structures that are related to the situation of each speaker. Formally, the conditional probability of a state for each utterance is a combination of the transition probabilities from the preceding state and a probability distribution over structures for each speaker.

Given a corpus of conversations made up of a sequence of posts, each of which is sequence of sentences, CSM automatically identifies structure language models and content topics. Structure language models are assumed to be expressed at the sentence level, that is, every sentence within a post is assumed to have one structure. CSM also assigns a content topic to each post.

Method

Figure 1 shows the pipeline of our method. We use unsupervised machine learning algorithms with human reflection to first induce a story schema, then fit the story schema to user post histories to form user profiles, and then use the profiles to build user trajectories. The unsupervised modeling provides a portal into the data that a human can then use to aid in interpretation. Our method interleaves unsupervised modeling with small amounts of human effort along the way to produce the final results. More specifically:

1. **Schema Induction and User Profile Formation:** We first use the CSM model to construct a prototype story schema from which to form user profiles. As part of this process, we perform Principle Component Analysis (PCA) (Jolliffe 2011) along with vectorization on text extracted using the schema.
2. **User Types Identification:** From the resulting vector-based representation, we perform clustering in order to assign users to different groups based on the text that instantiates the story schema when it is fit to the text from their post history. We refer to these clusters as user types.
3. **User Trajectory Transition Generation:** Next we segment user post histories and fit the schema to each segment in order to extract elements of a user's history that are salient at each time point. We then cluster these time point representations using the same method used to construct user types. In this case, the clusters represent user states. Last but not least we identify user trajectory transitions between user states for different types of users.

The remainder of the section introduces our dataset and explains each step in more detail.

Subreddits	# of users	# of posts
/r/MakeupAddiction	61,060	134,758
/r/SkincareAddiction	71,422	138,743
Total	121,050	273,501

Table 1: Dataset Statistics

Dataset

We select two specific subreddits to collect our dataset from, */r/MakeupAddiction* and */r/SkincareAddiction*. They represent the cosmetic domain well because they cover the most important topics, namely makeup and skincare, and they are the most cosmetic-related active subreddits with close to one million subscribers each. We collected all posts from both subreddits since their inception until September 2018. Table 1 shows the statistics of our dataset. Notice that there are users who post in both subreddits.

Table 2 gives an example of what user posts look like on Reddit. From this table we can tell that users share personal information such as experiences and stories in their posts.

In our work we choose the cosmetic domain as our main focus, but our method does not depend on anything that is specific to this domain. The cosmetic domain is interesting and suitable for generating user profiles for the following reasons:

- Topics covered in this domain are rich, spanning various aspects such as experience sharing, storytelling and etc. All this information is self-reported, and the source can be uniquely identified, so we can determine which posts are from the same user.
- The forums are popular, so the flow of new messages is high, which ensures we are able to collect a sufficiently large dataset.

User	Post
User1	I have acne oily skin. I was thinking of trying the say yes to tomatoes line. My skin isn't sensitive so I'm kinda looking forward to switching from my cetaphile to say yes to tomatoes. Or if I could get any other advice for skin care that'd be amazing!
User2	I wear sunscreen on my face everyday but my face is still always tanner than the rest of my body especially in the summer. It seems like most people's faces are lighter. Is there any tips I should know what causes this or what can I do to prevent it? Thanks!
User3	I have always had minor acne on my face but never suffered from PD. I recently broke out badly near my nose which then spread a little towards my eyes and lips area. I went to a dermatologist and she first said it's PD and prescribed me some gel. I am a week into the regimen and it seems to be getting worse. I asked the doctor again and she said to continue with the protocol and she now says that I have acne not PD and I should switch differin with Clindamycin. I have stopped using all creams and moisturizers (cerave) for the time being and want to see how my skin does without these. Curious to know what worked for everyone here and if I should see a different dermatologist.

Table 2: User Posts on Reddit

- To the best of our knowledge, no studies have been conducted on the cosmetic domain from Reddit posts, so the experiment has the potential to yield new knowledge.

After we extracted the posts from the two target subforums, the next step is to preprocess the data to prepare it for modeling. Note that we remove user "Automoderator" and "deleted" posts since they are not associated with any specific user. Text preprocessing includes the following three steps:

- Remove all URLs and punctuation.
- Lowercase all posts.
- Tokenize all posts.

In order to reduce issues with noise and sparsity at the modeling phase, we only include posts that have at least 3 sentences.

Story Schema Induction & User Profile Formation

The first step of inducing the story schema is to identify the typical sequences of sentences structures, which will then form an alphabet of schema elements in user posts from which a story schema will then be constructed. Specifically, we utilize CSM to identify a set of sentence structures underlying the corpus and then use PCA to identify the set of most strongly co-occurring structures (i.e., those that load onto the same principal component). These sets of structures then become the schema elements within constructed story schemas.

Structure	Example Post
Skin Situation	i think some background about my skin type and care is relevant i have oily skin even during dry winter months
Routine	my routine right now consists of natural drugstore facewashes and moisturizers for sensitive and acne prone skin using both in the am and the pm
Reviews	i have been using bobbi browns skin foundation and love the color but hate that it feels greasy
Outlook	what im looking for is basically more stable higher quality versions of the products i already have particularly eyeshadows and eyeliners

Table 3: Story Schema. Extracted by CSM and PCA

More concretely, to run CSM on user posts, we treat each post as analogous to what was a "conversation" in the original CSM experiments. For our data, each sentence in a post is treated as an "utterance" that consists of one sentence. So a discussion on Reddit with multiple posts is actually treated by CSM as a series of "conversations". Although different from how the model was used in the original paper, this setting makes sense because we are interested in the time-series of elements within each post to fit the schema. We found the optimal numbers of structure language models, of content topics, and of states to be 10, 5, and 5, respectively, to provide clear interpretation. The rationales and details of parameter values are described in *Appendix*.

Once structures of sentences are identified by CSM, we run PCA to select sets of co-occurring structures (within posts) to form story schemas that give us the latent structures of user posts. Based on the PCA analysis, we were able to extract three sets of co-occurring structures, out of which one set that consists of four structures forms a story whereas the other two made of two structures only form fractions of a story. Thus we decided to use the four-structure set as the story schema to use for illustration purposes in this paper. Table 3 gives an example for each of the structures in the schema learned by CSM. The sequence of story elements captured by the schema is this: First a user gives the background of their *skin situation*. Then they introduce their skincare or makeup *routine*. Next they describe their experiences in skincare or makeup, which mostly are product *reviews*. Last but not least they present their *outlook* by asking questions or seeking for opinions. After the schema is induced, we fit it on entire user post histories to extract sentences that are parts of the story to form user profiles. An example of extracted user profiles is shown in the *Result* section.

User Type Identification

Once we extract the story schema and use it to construct a user profile for each user, we move on to identify kinds of users based on their profiles. In order to find these types, we first represent each user by a vector through the following steps:

- We generate an utterance by term matrix. Each row represents a sentence and each column represents a word.
- We run PCA on the matrix and extract the top five principle components (PC). Therefore each sentence is represented by a 5-dimensional vector.
- For each user, we collect all the sentences for each element from the schema, average the values of each PC. Now each element for the user is represented by a 5-dimensional vector.
- We concatenate the vectors from all elements and represent each user by the resulting 20-dimensional vector.

Our goal is to cluster user vectors to find types of users. We use the K-means (Lloyd 1982) clustering algorithm to achieve this goal. Various Ks were tested through our experiments and the final K value was set to three for the best interpretation. Once the clusters are identified, we interpret the characteristics of each type of users by analyzing their profiles. The types of users and their characteristics are shown in the *Results* section.

User Trajectory Transition Construction

After we identify the types of users and interpret their characteristics, the next step is to extract salient user posts within the schema, i.e., remove idiosyncratic posts, segment the remaining set of posts into smaller units based on time periods, fit the posts within time periods to the story schema, and then cluster the instantiated schemas in order to identify types of user states. We extract salient posts by using the content topic of each sentence in a post, which is provided by CSM. All elements in the schema are mostly associated with one specific content topic so we filter out sentences that do not share the same one.

Once we extract salient posts, the next step is to segment posts into time units and cluster them to identify different user states. For segmentation, we observed that most users post a couple of times in a year and they may come back later, and the majority of users have either one or two active posting years. Therefore we decided to segment posts in a yearly manner (from 2013 to 2018). For clustering, we again perform vectorization, PCA and K-means clustering to achieve this goal. Various Ks were tested through our experiments and the final K value was set to three based on observed interpretability of the resulting clusters. We then interpret the characteristics associated with clusters by reading posts that are close to each cluster's centroid.

After we identify types of user states, the final step is to construct representations of user trajectories from the user state transitions identified for each user type. Based on the distribution over observed sequences of user states, we then create a transition diagram to show how states progress from one to the other for each user type. The types of user states and the transition diagrams are displayed in the *Results* section.

Type	Characteristic	%
Type 1	dry skin	37
Type 2	acne prone skin	8
Type 3	oily skin	55

Table 4: User Types. From clustering story schemas fit to entire user post histories

Results

In this section we first present types of users and types of user states identified. Then we show user trajectory transition diagrams along with analysis. Last but not least we demonstrate how we utilize user profiles as a lens to examine users' product experiences.

User Types & User States

Table 4 shows the user types and their percentages. From this table we can see that there are three types of users posting in these subreddits and the users are characterized by their skin types. We have more than half of the users with oily skin, slightly more than one third of the users with dry skin and the rest with acne prone skin. Generally speaking, people with dry skin tend to use hydrating products and look for products to reduce the dryness. People with acne prone skin tend to report acne treatments in their routine and seek better solutions or a combination of products to treat their acne. People with oily skin tend to have a form of deep cleanse in their routine.

We also give an example of an acne prone user's post snippets based on the story schema to demonstrate what a user profile looks like for each user type. Note that *am* stands for morning and *pm* stands for evening or night in the routine. Both *epiduo* and *accutane* are acne treatments:

...i have acne all over my face and redness...
(Skin Situation)

...my current routine am cetaphil face wash pm cetaphil face wash epiduo once per week accutane 20mg...
(Routine)

...anything that touched it would make it red and it burns...
(Review)

...is moisturizer going to help or hinder my acne...
(Outlook)

Table 5 shows different types of user states and their percentages. There are three user states in total. A little more than half of the user states are product focused, slightly more than one third of the user states are problem focused, and the rest describe users' respective acne journey specifically. Generally speaking users either have skin related problems and try to find help, or share their knowledge and experiences with certain products. However it is worth mentioning that the acne problem stands out from all the other problems,

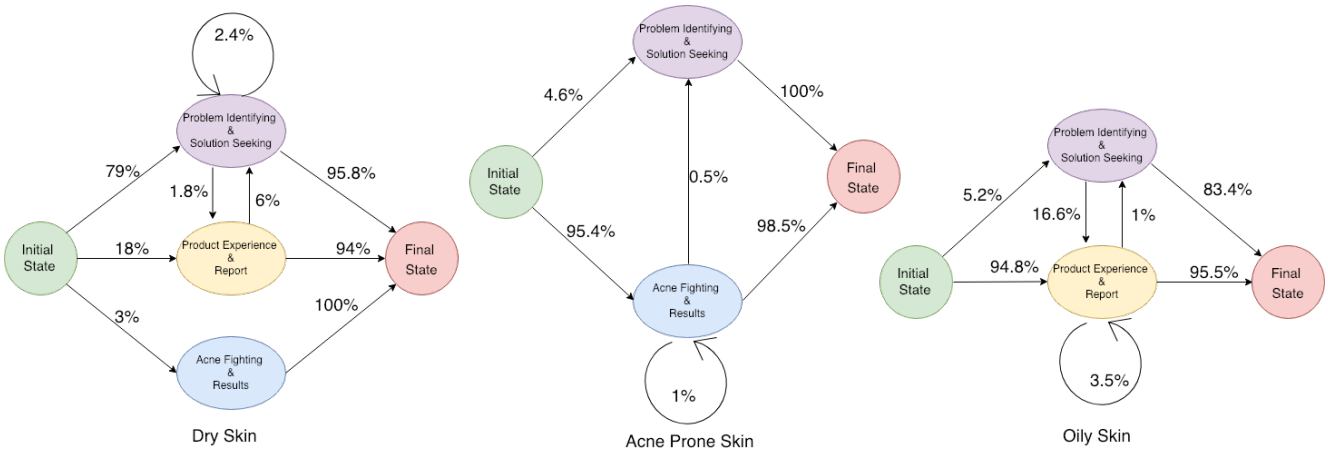

Figure 2: User Trajectory Transition Diagrams. Initial states are represented in green circles. Final States are represented in red circles. Different types of user states are represented in purple, yellow and blue circles respectively.

State	Characteristics	%
State 1	Problem Identifying & Solution Seeking	36
State 2	Product Experience & Report	55
State 3	Acne Fighting & Results	9

Table 5: User States. From clustering story schemas fit to year long intervals of time within user post histories

which illustrates that it is a more common issue that happens to users on these subreddits.

Here we show an example of snippets of user posts for each user state:

...my tap water comes up between a 910 ph and its wreckling my skin is there anything i can do...
(Problem Identifying & Solution Seeking)

...i put innisfree green tea balancing lotion aloe vera gel rosehip oil and vaseline on my face in that order over night and my skin is already so much more hydrated looking than it was...
(Product Experience & Report)

...struggled with cystic acne for 34 years found /r/skincareaddiction and face finally cleared up...
(Acne Fighting & Results)

User Trajectory Transition Diagram

Figure 2 shows the user trajectory transition diagrams for each user type. The initial state and final state are represented on the sides in each diagram and the other ones are pictured in the middle. An arrow represents a transition from one state to another along with its probability on top of the the arrow.

From this diagram we can tell that across the board, the majority of users only have one active posting time point, i.e., they only post in one year and do not come back. Dry skin users occupy all the user states whereas acne prone users and oily skin users only involve two out of the three user states, which shows that comparing to the other two skin types, the dry skin type generally has more diverse problems and experiences. Specifically, dry skin users more frequently start with asking for advice, then they come back later to report on their experiences. This suggests that they come into the forum at an earlier stage in their struggle with their condition. They are also more likely to return to an advice seeking state after having tried something. This indicates that things work for a while and then stop working, so they have to try something new.

Among the three skin types, acne prone skin users most often tell their stories – what they tried and what results they got – as an information sharing activity rather than coming in to ask advice, though some users start by asking for advice. It is interesting to see from the diagram that dry skin and acne prone skin users are more focused on problem solutions in general, whereas oily skin users are more focused on products. In addition, the acne problem does not only happen to the acne prone users, but sometimes to dry skin users as well. On the other hand oily skin users do not seem to report that problem. Furthermore, some acne prone users tend to have re-occurring acne problem but only once for other skin problems. In contrary, dry skin users seem to only deal with acne once but other issues more than once.

Oily skin users most frequently start talking about their experiences with certain products. They often mention their problems as well, but they are more product-oriented in comparison to dry skin users, who elaborate more on their problems. Both dry skin and oily skin users tend to transition between talking about problem-related experiences and product-related experiences. However again, dry skin users focus on the problem side and oily skin users focus on the product side.

Table 6 gives some specific transition examples from users that have two active years of posting. The dry skin user first reported their dryness issue and asked for opinions on a solution. The next time when they posted they typically

User Type	User State 1	User State 2
Dry Skin	**Problem** ...my skin tends to be dry... ...i clean my face twice a day with... ...do you think its better to get a lotion that already includes spf in it or to buy them separately...	**Product** ...i choose this moisturizer because of the niacinamide...
Acne Prone Skin	**Acne** ...i have acne on my cheeks now which ive never had before i attribute this to being off birth control and poor diet both of which im working on...	**Acne** ...with a focus on moisture and acne control...
Oily Skin	**Product** ...at the time i had been using neutrogena ultra sheer liquid sunscreen and after reading up a bit on different types of sunscreens i decided to ditch it for a physical...	**Product** ...i tried it on my face and it was kind of meh but when i rubbed some of the extra essence into my hands it was like a miracle...

Table 6: User Trajectory Transition Examples

Year	Skin Type	Routine	Review	Outlook
2015	dry skin	**A** milk cleanser **B** day cream moisturizer	burning sensation and redness after moisturizer	looking to try **C**
2016	dry skin	**C** milk cleanser and moistuizer oil cleansing	oil cleansing is better for hydrating skin comparing to milk cleanser at first burning sensation and redness later so had to stop	why the burning sensation
2017	dry skin	**D** cleanser and moisturizer	oil in **D** product cost redness and breakout	non redness and hydrating moisturizer

Table 7: Product Experience Examination Example. Real brand names are substitued by A, B, C and D.

talk about a specific product they used. The acne prone user described their acne fighting experiences in both years when they posted. The oily skin user reported a specific product they used both times when they made posts.

User Profiles

We fit the story schema to full user post histories to construct user profiles, which provide us with contextualized information and enables us to better understand users' product experiences over time. This information otherwise would not be available just from a single product review out of context. We illustrate how we use profiles as a lens to examine user experiences with specific products by an example.

Table 7 summarizes a user profile conducted by applying the story schema to their full post history from 2015 to 2017. From this profile we can tell that this dry skin user was having the burning sensation and redness problem with their skincare products, particularly a moisturizer, and was trying various products during the time they posted. Specifically they tried different cleansers and moisturizers from brand A, B, C and D. They first tried A and B but these brands seemed to cost a burning sensation and skin to become red. Then the user moved on to C along with oil cleansing treatment. However, the user had the burning sensation again after treating with oil. Next the user tried D but reported that the oil ingredient in D made the skin to break out and turn red. In the end the user was looking for a hydrating moisturizer that does not cost redness. According to this user profile, first we could infer that oil was probably what has been causing the issue from the beginning since the burning sensation and redness seemed to happen whenever oil was involved. In addition, we could speculate that brand A and B products contain oil in them. Furthermore, since we know what brands and products they already tried, the reason to their problem i.e., the oil ingredient, and the user's skin type, i.e., dry skin, we could recommend this user to either go back to brand C or try some new brands targeting the dry skin type that do not include oil. Without the profile which provides

us the characteristic of this user and their complete experience with different products, we would not be able to figure out what was causing the issue and what products would be suitable for this user.

Conclusion & Future Work

In this paper, we have presented a pipeline including a novel human-in-the-loop computational pipeline for the purpose of extraction of story schemas and construction of user profiles. The user profiles themselves are structured representations that can be used as a lens for understanding user trajectories over time, and how the trajectories are associated with characteristics of users. Our approach gathers a comprehensive user history from the perspective of selected subreddits, i.e., user experiences over time, which enables us to perform context-level analysis to better understand a user's or a type of users' preferences and needs. Our method takes advantages of both unsupervised modeling approach and human interpretation, and provides a new perspective in constructing and interpreting user profiles. We establish a story schema that captures the typical elements of user experiences in the realm of cosmetics. Furthermore we build user profiles by fitting the story schema on their full post histories as well as their segmented post histories. In addition we identify types of users and types of user states to generate user trajectory transition diagrams, which illustrate how different types of users' experiences change over time and how trajectories are associated with different types of users. We are also able to utilize the constructed user profiles to examine product usage experiences in context.

One limitation of this work is that in this paper we describe the process and individual parts of creating user profiles and how we use them to interpret user trajectories. This leads to one possible future direction, which is to develop a more integrated model. Moreover, in the work reported here we only focus on the cosmetic domain, though our goal has been to lay the foundation for similar analyses across a plethora of topic areas. In the future it is possible to extend our method onto other domains, such as environmental focused subreddits, in order to adopt a broader perspective on user type and user state, or even other discussion platforms.

References

Barzilay, R., and Lee, L. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *arXiv preprint cs/0405039*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Ekstrand, M. D.; Riedl, J. T.; Konstan, J. A.; et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction* 4(2):81–173.

Farseev, A., and Chua, T.-S. 2017. Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *AAAI*, 95–101.

Fiesler, C.; Jiang, J. A.; McCann, J.; Frye, K.; and Brubaker, J. R. 2018. Reddit rules! characterizing an ecosystem of governance. In *ICWSM*, 72–81.

Gjurković, M., and Šnajder, J. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 87–97. Association for Computational Linguistics.

Griffiths, T. L.; Steyvers, M.; and Tenenbaum, J. B. 2007. Topics in semantic representation. *Psychological review* 114(2):211.

Jo, Y.; Yoder, M. M.; Jang, H.; and Rosé, C. P. 2017. Modeling dialogue acts with content word filtering and speaker preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2017, 2169. NIH Public Access.

Jolliffe, I. 2011. Principal component analysis. In *International encyclopedia of statistical science*. Springer. 1094–1096.

Lee, D.; Jeong, M.; Kim, K.; Ryu, S.; and Lee, G. G. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions on Audio, Speech, and Language Processing* 21(11):2451–2464.

Li, J.; Ritter, A.; and Hovy, E. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 165–174. Association for Computational Linguistics.

Liang, P.-W., and Dai, B.-R. 2013. Opinion mining on social media data. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 2, 91–96. IEEE.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Lloyd, S. 1982. Least squares quantization in pcm. *IEEE transactions on information theory* 28(2):129–137.

Mei, Q.; Ling, X.; Wondra, M.; Su, H.; and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, 171–180. ACM.

Mislove, A.; Viswanath, B.; Gummadi, K. P.; and Druschel, P. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, 251–260. ACM.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; Smith, N. A.; et al. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm* 11(122-129):1–2.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Paul, M. J. 2012. Mixed membership markov models for unsupervised conversation modeling. In *Proceedings of the*

2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 94–104. Association for Computational Linguistics.

Penalver-Martinez, I.; Garcia-Sanchez, F.; Valencia-Garcia, R.; Rodriguez-Garcia, M. A.; Moreno, V.; Fraga, A.; and Sanchez-Cervantes, J. L. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications* 41(13):5995–6008.

Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*, 1–7. Citeseer.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180. Los Angeles, California: Association for Computational Linguistics.

Schafer, J. B.; Frankowski, D.; Herlocker, J.; and Sen, S. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer. 291–324.

Tan, C. 2018. Tracing community genealogy: How new communities emerge from the old. *arXiv preprint arXiv:1804.01990*.

Wallace, B. C.; Trikalinos, T. A.; Laws, M. B.; Wilson, I. B.; and Charniak, E. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1765–1775.

Wang, J.; Li, S.; Jiang, M.; Wu, H.; and Zhou, G. 2018. Cross-media user profiling with joint textual and social user embedding. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1410–1420. Association for Computational Linguistics.

Appendix

Parameter Values

Various parameter values were tested and the final parameter setting was chosen based on model performance and the parameter setting suggested in the original paper (Jo et al. 2017).

We found the optimal number of structures to be 10. Higher numbers tend to capture too content-specific structures, and lower numbers too general structures. The optimal number of content topics is 5, which indicates that the corpus is focused on cosmetics and the content is relatively common across the corpus. The number of states reflects different patterns of structure composition within a post, and 5 states were found to be optimal. More states tend to capture too post-specific structures, and less states cannot account for the diversity of structures.

Parameter $\nu \in [0, 1]$ is the weight on state transition probabilities (as opposed to speaker preferences) for determining an utterance's state. 1 means only state transition probabilities are considered, and 0 means only speaker preferences are considered. In our study, a high weight on state transitions ($\nu = 0.9$) helped to find schemas that represent common profiles of users, probably because this setting drives the model to identify sentence structures that are less sensitive to individual users. Parameter $\eta \in [0, 1]$ is the weight on structure language models (as opposed to content topics) for generating words. 1 means that all words are generated from structure language models, and 0 means only from content topics. Our setting ($\eta = 0.8$) filters out 20% of words as content. This is quite a large proportion compared to the original paper, meaning that the corpus has a relatively clear separation between structures and content.

Other hyperparameters for the model were set identically to the original paper: $\alpha^F = \gamma^A = 0.1, \alpha^B = \gamma^S = 1, \beta = 0.001$.