Gender and Interest Targeting for Sponsored Post Advertising at Tumblr

Mihajlo Grbovic†, Vladan Radosavljevic†, Nemanja Djuric†,
Narayan Bhamidipati†, Ananth Nagarajan‡
†Yahoo Labs ‡Yahoo, Inc.
701 First Ave, Sunnyvale, CA, USA
{mihajlo, vladan, nemanja, narayanb, ananth}@yahoo-inc.com

ABSTRACT

As one of the leading platforms for creative content, Tumblr offers advertisers a unique way of creating brand identity. Advertisers can tell their story through images, animation, text, music, video, and more, and can promote that content by sponsoring it to appear as an advertisement in the users' live feeds. In this paper, we present a framework that enabled two of the key targeted advertising components for Tumblr, gender and interest targeting. We describe the main challenges encountered during the development of the framework, which include the creation of a ground truth for training gender prediction models, as well as mapping Tumblr content to a predefined interest taxonomy. For purposes of inferring user interests, we propose a novel semi-supervised neural language model for categorization of Tumblr content (i.e., post tags and post keywords). The model was trained on a large-scale data set consisting of 6.8 billion user posts, with a very limited amount of categorized keywords, and was shown to have superior performance over the baseline approaches. We successfully deployed gender and interest targeting capability in Yahoo production systems, delivering inference for users that covers more than 90% of daily activities on Tumblr. Online performance results indicate advantages of the proposed approach, where we observed 20% increase in user engagement with sponsored posts in comparison to untargeted campaigns.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data Mining

General Terms

Information Systems, Algorithms, Experimentation

Keywords

Data mining; computational advertising; audience modeling

KDD '15, August 10-13, 2015, Sydney, NSW, Australia

DOI: http://dx.doi.org/10.1145/2783258.2788616.

1. INTRODUCTION

In recent years, online social networks have evolved to become an important part of life for online users of all demographic and socio-economic backgrounds. They allow users to easily stay in touch with their friends and family, discuss everyday events, or share their interests with other users with the click of a button. Tumblr is one such social network, representing one of the most popular and fastest growing networks on the web. Hundreds of millions of people around the world come every month to Tumblr to find, follow, and share what they love. Consequently, the Tumblr network represents a gold mine of content, comprising around 200 million blogs on different topics such as travel, sports, or music, with 85 million user posts being published on a daily basis. This wealth of user-generated data opens a great opportunity for advertisers, allowing them to promote their products through high-quality targeting campaigns to both blog visitors and blog owners [19].

The prevalent form of advertising on Tumblr is through *sponsored posts* that appear alongside regular posts in the user's *dashboard*, the central page for a Tumblr user, displaying the most recent posts of followed blogs in the form of a stream. This form of advertising, where advertisements resemble native content in the stream, is often referred to as *native advertising*. Native advertisements are usually aesthetically beautiful and highly engaging, which typically makes them more enjoyable than regular display ads [4]. Tumblr launched its native advertising product in May of 2012. Since then, the number of advertisers (or brands) on the platform has grown steadily and reached a milestone of 100 advertisers in April of 2013. Moreover, 8 of the 10 most valuable brands are advertising on Tumblr[1], while sponsored posts have generated more than 3 billion paid ad impressions since the launch of the Tumblr advertising product[2]. However, a huge marketing potential of Tumblr [19] has not been fully exploited, due to the fact that targeting against specific interest and demographic audiences, a targeting component that Tumblr was missing, has become an industry standard and many advertisers are in need of such a solution.

Building interest targeting products on social and microblogging platforms is an important research topic, discussed previously by several researchers [14]. However, due to its distinct characteristics, Tumblr poses novel challenges, which we explain in detail in this paper. In particular, the

[1] `marketr.tumblr.com`, accessed June 2015
[2] `www.comscore.com`, accessed June 2015

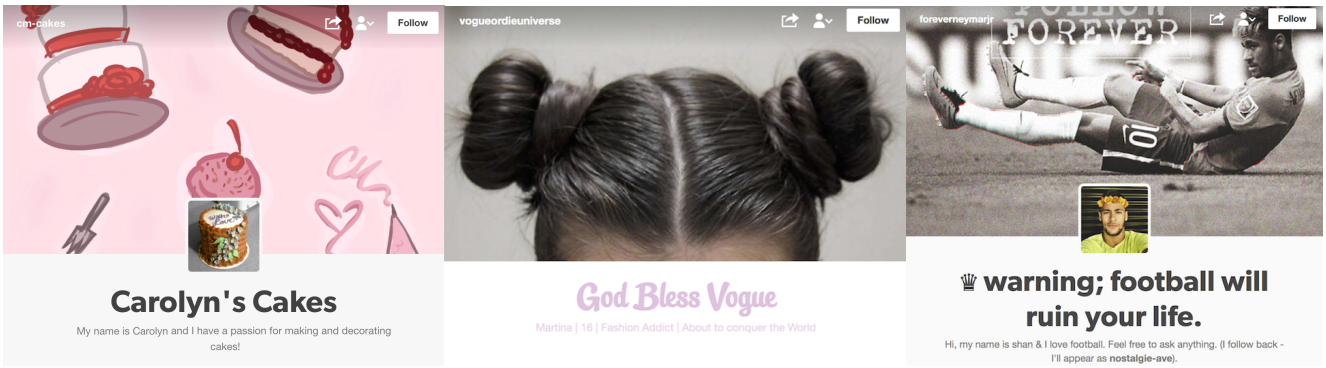

Figure 1: Examples of blog titles (larger, bolded font) and blog descriptions (smaller font)

content and language used on Tumblr have distinct characteristics that needed to be accounted for during the modeling. For instance, users often use tags to summarize the text in their posts. However, the language styles used in the tags and post text are different (e.g., the tag "hp" and the word "hp" found in posts have different meanings, "Harry Potter" and "Hewlett-Packard", respectively). Moreover, unlike the popular social platform Facebook, which contains a large amount of social interactions but a limited amount of content, or the microblogging platform Twitter, which contains an intermediate amount of social interaction and content, Tumblr represents a unique combination of a rich and diverse content platform and a dynamic social network. To make use of this vast advertising potential, in this paper we propose to classify user-generated Tumblr content into a standard multi-level *general-interest taxonomy*[3] that advertisers commonly use for defining their targeting campaigns, opening doors to high-quality audience segmentation and modeling for purposes of ad targeting. However, inferring categories of user posts is a challenging task, given huge quantities of unlabeled data being posted every day and very limited amount of labeled data, typically obtained by editorial efforts. To this end, we propose a novel semi-supervised neural language model, capable of jointly learning embeddings of post keywords, post tags, and category representations in a common feature space. The neural model was trained on a large-scale data set comprising 6.8 billion posts, with only a fraction of manually categorized content.

Targeting pipelines described in this paper are being used to show ads to millions of users daily, and have substantially improved Tumblr's business metrics following the launch. We detail our path to developing targeting capabilities for Tumblr, where one of the key steps was creating user profiles, based on users' activities that include publishing blog posts, following other blogs, or liking posts. Then, we describe how both demographic and interest predictive models were built based on the created profiles.

Lastly, we emphasize that the privacy of users is of critical importance to Yahoo. Therefore, we were constrained in regards to what data we can use. Specifically, user profiles were created solely from data which users share publicly with others, including contents of blog posts, blog titles and descriptions, as well as follow, like, and reblog actions. This data is publicly available through Tumblr Firehose data source[4]. Other user activities, such as user searches on Tumblr, which blogs they visited or where they clicked, are all considered to be sensitive data and were not used in any way for the development of the presented ad targeting models.

2. RELATED WORK

Personalization is defined as "the ability to proactively tailor products and product purchasing experiences to tastes of individual consumers based upon their personal and preference information" [7], and it has become an important and very lucrative topic in the recent years. Personalization of online content may lead to improved user experience and directly translate into financial gains for online businesses [17]. In addition, personalization fosters a stronger bond between customers and companies, and can help in increasing user loyalty and retention [2]. For these reasons it has been recognized as a strategic goal and is the focus of significant research efforts of major internet companies [8, 15].

We consider personalization through the prism of ad targeting [11], where the task is to find the best matching ads to be displayed for each individual user. This improves the user's online experience (as only relevant and interesting ads are shown) and can lead to increased revenue for the advertisers (as users are more likely to click on the ad and make a purchase). Due to its large impact and many open research questions, targeted advertising has garnered significant interest from the community, as witnessed by a large number of recent workshops[5] and publications [5, 9, 14].

One of the basic approaches in ad targeting is to target users with ads based on their demographics, such as age or gender. Historically, this method has proven to work better than targeting random users. However, while for some products this type of targeting may be sufficient (e.g., women's makeup, women's clothing, man's razors, man's clothing), for others it is not effective enough and more involved profiling of users is required. A popular targeting approach that addresses this issue is known as interest targeting, in which users are assigned interest categories based on their historical behavior, such as "sports" or "travel" [1]. Typically, a taxonomy is used to decide on the targeting categories, and a model is learned to categorize user activities and estimate their interest in each category. Interest targeting is known to

[3] `http://www.iab.net/QAGInitiative/overview/taxonomy`, accessed June 2015

[4] `gnip.com/sources/tumblr`, accessed June 2015

[5] `www.targetad-workshop.net`, accessed June 2015

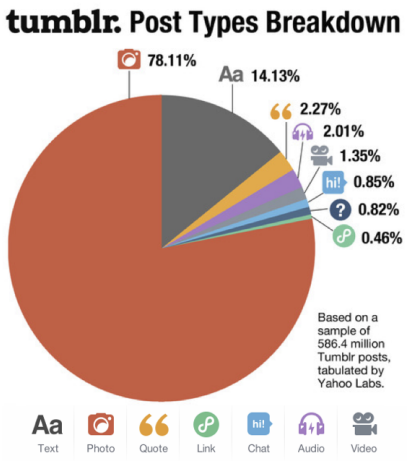

Figure 2: Distribution of Tumblr post types

Figure 3: Example of Tumblr blog post

build good brand awareness with relevant audience, which has already shown interest in the corresponding category. In this paper we follow this targeting approach. Alternatively, advertisers may be interested to go a step further and optimize for current intent as opposed to long-term interest of users, typically done by assigning categories to actual ads, and training a machine learning model to estimate the probability of an ad click in that category [10, 20]. For each ad category a separate predictive model can be trained and evaluated on the entire user population, with N users with the highest score selected for ad exposure.

To the best of our knowledge, the Tumblr social network has been considered by only a handful of scientific studies. In [3, 18] the problem of blog recommendation is discussed, while in [6] the authors explore social norms on the social network. However, our work is the first that addresses an important problem of ad targeting on Tumblr.

3. WHAT IS TUMBLR?

Tumblr[6] is one of the most popular social blogging platforms on the web today, where users can create and share posts with the followers of their blogs. According to the data from January 2015[7], there is a total of 221.6 million blogs on Tumblr, which jointly produced over 102.7 billion blog posts. With a large number of new users signing up every day, it is currently the fastest growing social platform[8].

3.1 User activities on Tumblr

To register a Tumblr account, a valid e-mail address is required, along with a primary username (which becomes a part of the blog URL) and a confirmation of age. Once created, a Tumblr blog contains a profile picture, blog title, and blog description appearing at the top (see Figure 1), followed by a stream of blog posts bellow. The first blog created by a registered user is considered their primary blog. In addition, very small portion of users maintains one or more secondary blogs. A Tumblr user is uniquely identified by a blog ID of their primary blog, and throughout the paper we will use terms "blog" and "user" interchangeably.

Common user activities of Tumblr users include the following actions: 1) creating a post on one's blog; 2) sharing a post created by another blog, called *reblogging* (a reblogged post will appear on the user's blog); 3) liking a post by another blog; and 4) following another blog. Similarly to Twitter, follow connections on Tumblr are unidirectional. However, unlike Twitter, users can create longer and richer content in a form of several post types, such as text, photo, quote, link, chat, audio, and video. The posts are shown in user's dashboard, ordered such that more recent posts appear closer to the top. The most popular types of blog posts are photo and text posts, which, based on the analysis published in [21], together cover more than 92% of all content on Tumblr (see Figure 2 for detailed distribution of post types). In addition, any post type can be annotated with words starting with the "#" sign (called *tags*) that concisely describe a post and allow for easier browsing and searching. Additional metadata that describes a post includes photo captions in photo posts, post titles in text posts, and artists names in audio posts. An example photo post is shown in Figure 3. Tags are displayed bellow the photo caption (e.g., *#gadgets* and *#tech*), while buttons for reblog and like actions are located in the bottom right corner.

3.2 Advertising on Tumblr

Advertising on Tumblr is implemented through the mechanism of sponsored (or promoted) posts shown in user's dashboard. This is similar to how advertising works on Twitter and Facebook. A sponsored post can be a video, an image, or simply a textual post containing an advertising message. In Figure 4, we show an example of a sponsored post and how it appears on desktop and mobile dashboards. Similarly to organic (or non-promoted) posts, sponsored posts can propagate from user to user in the network by means of reblogs, and users can also "like" the promoted post. Both likes and reblogs can be seen as an implicit form of acceptance or endorsement of the advertising message. Moreover, just like other posts, sponsored posts are supplemented with notes on who liked and reblogged them.

[6] www.tumblr.com, accessed June 2015
[7] www.tumblr.com/about, accessed June 2015
[8] http://t.co/3txHFRJreJ, accessed June 2015

Figure 4: Example of Tumblr sponsored post

Interestingly, while user-generated, organic posts are reblogged 14 times on average, sponsored posts are reblogged 10,000 times on average[9]. We have observed that 40% of engagements with sponsored posts are reblogs, likes, and follows. Moreover, every four reblogs of a sponsored post result in 6 downstream reblogs from followers, leading to content longevity, while one third of reblogs of sponsored posts are present for 30 days or more after the initial post.

4. TUMBLR DATA

In this section we describe data sets comprising user activities and post contents, which were utilized to create user profiles. In particular, user activities include actions such as posts, likes, follows, and reblogs, while post contents include tags, title and body for text posts, artist names from audio posts, as well as tags and captions for photo posts.

4.1 Used data sets

Once signed into Tumblr, a user can follow other users' blogs. The follow action is one-directional as it does not require the followed user to follow back. For the purpose of this study, we extracted a sub-graph which contained 96.9 million unique nodes (i.e., users) and 5.1 billion edges (i.e., follows), out of which 36.4 million are bidirectional (18.2 million pairs of users that follow each other). The data set included more than 26.1 billion activities on Tumblr. As mentioned earlier, an entire activity log is publicly available through a data feed called Firehose.

To create user profiles for targeting, textual contents of all posts were collected, including photo captions, tags, titles, and bodies. In addition, every time a user performs a post or reblog activity, Firehose lists the user's blog title and blog description, which were also employed to represent a user. As we can see in Figure 1, a blog title and description often provide useful information with respect to targeting, such as the user's first name, age, and even declared interests (e.g., statements such as "fashion addict" or "I love football").

[9] http://yhoo.it/1vFfIAc, accessed June 2015

Table 1: User data extracted from Tumblr Firehose

Declared	Content	Actions
blog title	post tags	reblog
blog description	photo captions	like
	text post title	follow
	text post body	
	audio post artists	

4.2 Data processing

In order to obtain useful representations of user profiles, we propose to extract keywords from available blog information, which requires data preparation and processing. Given the extracted blog data, including title, content, and tags, we first removed all HTML tags, followed by the extraction of bigrams and the removal of common English stopwords.

In particular, it is common for certain words to appear together more often than some others (e.g., words "credit" and "card"), and we aim to capture those bigrams and use them in keyword-based user profiles. To detect bigrams, we use a procedure that counts the unigram and bigram appearances, and for each combination of words w_i and w_j calculates the following score,

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i \text{ and } w_j \text{ together})}{\text{count}(w_i)\ \text{count}(w_j)}. \qquad (4.1)$$

Finally, bigrams with a score above a certain threshold were extracted from the blog contents, along with the remaining unigrams. On the other hand, post tags are originally formed as n-grams by the users (e.g., *#chess rules*), and were extracted in their original form.

4.3 User profiles

Available data sources were used to create user profiles. In particular, we extracted three distinct groups of user-related data: 1) declared; 2) content of posts; and 3) actions. The specific components included in each of the data groups are listed in Table 1. From each group we extracted features to represent the users as described below.

Declared data consists of information provided during sign-up, including keywords from the blog title and blog description extracted using the method described in Section 4.2. We counted the keyword frequency in a user's blog title and description, and stored the counts along with a timestamp of the latest log-in as a part of user's profile.

Content features were formed from the textual contents of posts which a user either created or reblogged. The main content feature types include: 1) post tags; 2) keywords from the post title and body; 3) keywords from the captions of photo post; and 4) artist names from audio posts. In this way we collected several millions of distinct keywords that were used to obtain rich representation of user profiles. To illustrate content keyword extraction from the Firehose, consider that user u_i at timestamp t used tag *#hp* five times and tag *#nba* eight times, keyword *football* two times in post titles, and posted ten times an audio post with a song from artist *Shakira*. Then, the resulting user profile would be $u_i = \{tag : \{\#hp, t : 5; \#nba, t : 8\}, title : \{football, t : 2\}, artist : \{shakira, t : 10\}\}$.

Action features include follows, likes, and reblogs. If user u_i follows user u_j at timestamp t, we create an indicator feature $follows : \{j, t : 1\}$ and add it to the u_i user's profile. Similarly, if user u_i likes user's post, we create a feature that

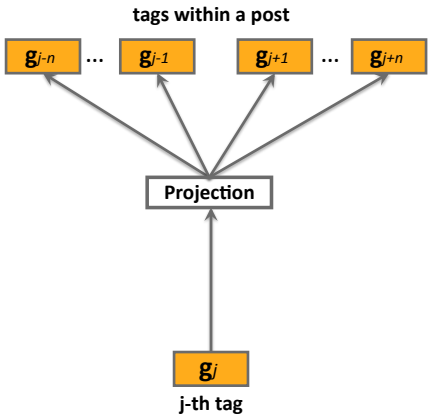

Figure 5: Unsupervised skip-gram model

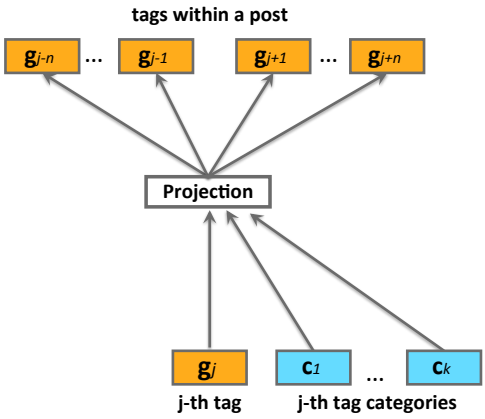

Figure 6: Semi-supervised skip-gram model

keeps record of the number of likes m, as $likes : \{j, t : m\}$, and update the user's profile accordingly.

The timestamps used in feature engineering represent the day on which the activity happened. For the experiments presented in this paper we subsampled Tumblr users to obtain 80 million user profiles. The total number of unique features was 1.4 million, and on average a user had 380 non-zero features.

5. INTEREST PREDICTION

The goal of our work is to identify user groups with interests in certain topics, such as music, travel, cooking, or books, in order to allow advertisers to target segmented Tumblr audiences, as well as to infer user demographics (discussed in Section 6). As the topic interests may be defined at various levels of granularity, to avoid sparsity problems while still providing useful and actionable interest categories, user interests are often classified into a pre-determined hierarchical interest taxonomy that the advertisers commonly use. However, to be able to create effective user interest classifiers, a modeler requires a sufficient amount of labeled data. Yet, for the problem of the scale of Tumblr interest prediction, this can be a daunting task for human editors. For that reason we propose to use a novel semi-supervised classification approach [12] based on the recently proposed word2vec model [16], which efficiently and seamlessly makes use of large amounts of unlabeled and a limited amount of labeled data for learning effective content classifiers.

5.1 User interest taxonomy

We decided to classify keywords into the General Interest Taxonomy (GIT), used by the Yahoo Gemini advertising platform for native advertising[10]. The GIT is carefully derived based on Interactive Advertising Bureau (IAB) taxonomy recommendations, in order to meet advertiser needs and protect Yahoo's interests. The GIT has a two-level hierarchical structure, such that advertisers can adjust the audience reach by utilizing broader or narrower interest categories. The top level of the taxonomy contains 23 nodes (e.g., "Automotive", "Pets", "Travel"), while the second level contains 130 nodes which represent more focused interests (e.g., "Automotive/SUV", "Automotive/Luxury", "Pets/Dogs").

[10]`http://gemini.yahoo.com`, accessed June 2015

5.2 Semi-supervised classification

In this section, we describe a recently proposed classification approach [12] based on the skip-gram model [16], which is used to categorize keywords into the GIT taxonomy. For conciseness, we describe the proposed model on the assumption that it is applied to tag categorization. However, it is straightforward to use the same methodology for categorization of keywords originating from blog titles and descriptions, as well as from text, audio, and image posts. Thus, we consider the task of tag classification, where the goal is to classify tags into one or more interest categories. In order to address this problem, we learn tag representation in a low-dimensional vector space using neural language models that are applied to historical Tumblr posts.

More specifically, let us assume that we are given N posts. In the post logs found in Firehose, each post p is recorded along with the tags g_j, $j = 1...M$, where M represents the number of tags in the post. Given the data set $\mathcal{D}$ of all posts, the objective is to find a vector representation of tags in which semantically similar tags are nearby in the vector space. For this purpose, we extend ideas originating from recently proposed language models, as described in the remainder of this section.

The skip-gram (SG) model involves learning representations of tags in a low-dimensional space from post logs in an unsupervised fashion, by using the notion of a blog post as a "sentence" and the tags within the post as "words", borrowing the terminology from the Natural Language Processing (NLP) domain (see Figure 5). Tag representations using the skip-gram model [16] are learned by maximizing the objective function over the entire $\mathcal{D}$ set of blog posts defined as follows,

$$\mathcal{L} = \sum_{p \in \mathcal{D}} \sum_{g_j \in p} \sum_{-n \leq m \leq n, m \neq 0} \log \mathbb{P}(g_{j+m}|g_j). \quad (5.1)$$

Probability $\mathbb{P}(g_{j+m}|g_j)$ of observing a neighboring tag g_{j+m} given the current tag g_j is defined using the soft-max,

$$\mathbb{P}(g_{j+m}|g_j) = \frac{\exp(\mathbf{v}_{g_j}^\top \mathbf{v}'_{g_{j+m}})}{\sum_{k=1}^{G} \exp(\mathbf{v}_{g_j}^\top \mathbf{v}'_k)}, \quad (5.2)$$

where $\mathbf{v}_g$ and $\mathbf{v}'_g$ are the input and output vector representations of tag g of user-specified dimensionality d, n defines the length of the context for tag sequences, and G is the

Table 2: Examples of categorized tags

Tag	Category
music	Arts & Entertainment/Music
fashion	Style & Fashion
song	Arts & Entertainment/Music
art	Arts & Entertainment
disney	Arts & Entertainment/Movies
style	Style & Fashion
photography	Hobbies/Photography
teen wolf	Arts & Entertainment/TV
food	Food & Drink

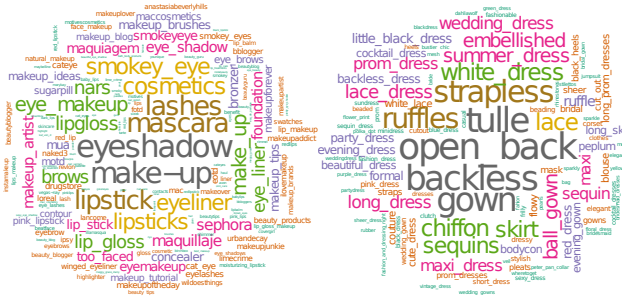

Figure 7: Nearest neighbors of tags: a) *#makeup*; b) *#dress*

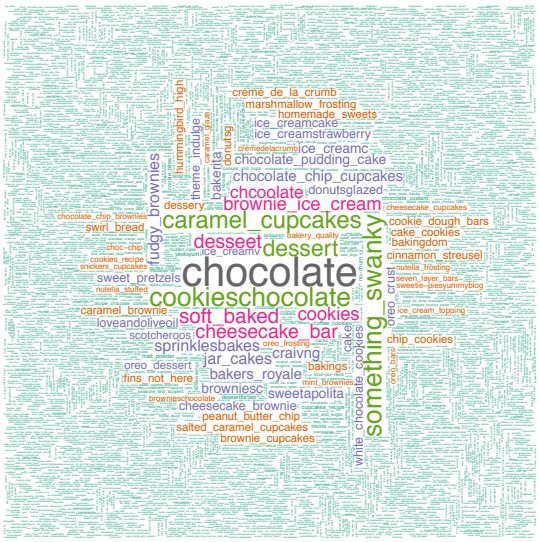

Figure 8: Nearest tags to "Food & Drink/Desserts"

number of unique tags in the vocabulary. Following training of the skip-gram model, tags that co-occur often and tags with similar contexts (i.e., with similar neighboring tags) will have similar vector representations.

The semi-supervised skip-gram (SS-SG) model assumes that some tags are labeled with categories from the GIT taxonomy. Then, we introduce a dummy category vector for each node of the taxonomy, and leverage the tag contexts in blog posts to jointly learn tag vectors and category vectors in the same feature space [12]. Given such setup, after learning the representations every tag from the vocabulary can be categorized by simply looking up the closest category vector in the joint embedding space.

Given a set of categorized tags, we extend $\mathcal{D}$ to obtain data set $\mathcal{D}_{ss}$ where available categories are imputed into post "sentences" p. In particular, labeled tags are accompanied by assigned categories, and every time a vector of a labeled central tag t_j is updated to predict the surrounding tags, vectors of categories assigned to g_j are updated as well. More formally, assuming the central tag g_j is labeled with C_j of C categories in total, $\zeta_j = \{c_1, \ldots, c_{C_j}\}$, the semi-supervised skip-gram learns tag and category representations by maximizing the following objective function,

$$\sum_{p\in\mathcal{D}_{ss}}\sum_{g_j\in p}\sum_{-n\le m\le n, m\neq 0}\Big(\log\mathbb{P}(g_{j+m}|g_j)+\sum_{c\in\zeta_j}\log\mathbb{P}(g_{j+m}|c)\Big). \tag{5.3}$$

Probability $\mathbb{P}(g_{j+m}|c)$ of observing tag g_{j+m}, given label c of the current tag g_j, is defined using the soft-max,

$$\mathbb{P}(g_{j+m}|c) = \frac{\exp(\mathbf{v}_c^\top \mathbf{v}'_{g_{j+m}})}{\sum_{k=1}^{G}\exp(\mathbf{v}_c^\top \mathbf{v}'_k)}. \tag{5.4}$$

This procedure allows us to seamlessly incorporate labeled and unlabeled data, and learn tag and category vectors in the common embedding space. Then, classification of tags amounts to a simple nearest-neighbor search among the category vectors. In Figure 6 we show graphical representation of the semi-supervised skip-gram model.

5.2.1 Training

The models are optimized using stochastic gradient ascent, suitable for large-scale problems. However, computation of gradients $\nabla\mathcal{L}$ in (5.1) and (5.3) is proportional to the vocabulary size G, which may be expensive in practice as G could easily reach several million tags. As an alternative, we used a negative sampling approach [16], which significantly reduces the computational complexity.

The data set used during the model training comprised 6.8 billion posts that contained tags. To collect category labels for some of the tags, we sorted the tags in a decreasing order of popularity and through editorial efforts labeled the top ones with one or more categories. This resulted in a total of 8,400 categorized tags used during semi-supervised training. We show examples of categorized tags in Table 2.

The representations were trained using a machine with 96GB of RAM memory and 24 cores. Dimensionality of the embedding space was set to $d = 300$, and the neighborhood size was set to $n = 5$. Finally, we used 10 negative samples in each vector update. Similarly to [16], most frequent tags were subsampled during training.

5.2.2 Inference

When the vector representations of all tags are learned, we can find similar tags for a given tag by straightforward k-nearest neighbor (k-NN) searches in the representation space. We use cosine distance [16] as a measure of similarity. To illustrate the usefulness of our approach, word clouds of neighboring tags to the tags *#makeup* and *#dress* are shown in Figure 7, where we see that semantically similar tags are grouped in same parts of the embedding space.

Similarly, we can find the most likely category for any tag by finding the nearest category in the vector space. To produce a high-confidence set of categorized tags, we retrieved only tags with a cosine distance of 0.7 or higher to the corresponding category vectors. This threshold was obtained

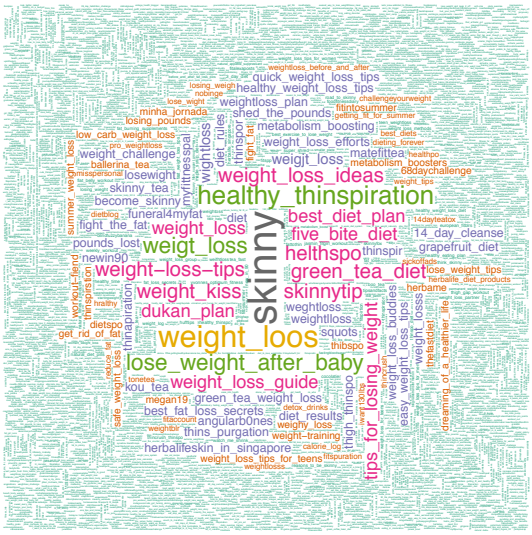

Figure 9: Nearest tags to "Health & Fitness/Weight Loss"

Table 3: Precision and recall of the competing methods

Method	Precision	Recall
LR-SG	0.71	0.65
k-NN-SG	0.82	0.62
SS-SG	**0.85**	**0.63**

through editorial evaluation of the results. In total, more than 380,000 tags were categorized into one or more categories with high confidence. To illustrate the quality of obtained classifiers, in Figures 8 and 9 we give word clouds of categorized tags for "Food & Drink/Desserts" and "Health & Fitness/Weight Loss" categories, respectively. A demonstration video of our tag categorization tool is available online[11].

5.2.3 *Evaluation*

In order to quantify the benefits of our approach, we trained method by excluding random 2,000 tags from the editorially labeled set, and evaluating on the held-out set by predicting the top-level GIT categories. We compared the SS-SG classification to the state-of-the-art logistic regression (LR) and k-NN methods, trained on the vectors learned by the original SG model. For LR classification (we refer to the method as LR-SG) we trained one classifier per interest category, while for k-NN (we refer to the method as k-NN-SG) for each held-out tag we found $K = 50$ nearest categorized neighbors and predicted its category by a majority vote. We report the results following a 5-fold cross-validation in Table 3. The results indicate that classification based on the proposed approach achieves higher precision than the competing methods, while at the same time maintaining competitive recall measure.

5.2.4 *Blog- and post-specific models*

To be able to map more of Tumblr content to the GIT taxonomy, we trained two additional SS-SG models: 1) for keywords from post titles and bodies; and 2) for keywords from blog titles and descriptions. To train the models, we followed a similar procedure as before. Editors provided 4,700 categorized keywords, which were used to form training data sets for SS-SG model learning. Post keyword vectors were trained using a data set comprising $N = 6.8$ billion posts, while blog title and description keyword vectors were trained using $N = 37.1$ million blogs. Then, to find high-confidence keywords for each category, we calculated cosine distances between learned vectors of categories and of keywords from the vocabulary. We retrieved 184,000 text blog keywords with a cosine distance of 0.7 or higher. We repeated the same procedure for keywords from blog titles and descriptions, resulting in 173,000 categorized keywords.

Note that the three models were trained separately due to language differences between these domains (i.e., between post tags, post text, and blog title and description text). To justify our claim, in Table 4 we show the nearest neighbors of the tag "hp" from the SS-SG model trained on tags and the word "hp" from the SS-SG model trained on post text. As we can see in the table, "hp" has different meanings in post tags and post text domains, referring to "Harry Potter" and "Hewlett-Packard", respectively.

Table 4: Language differences in post tags and post text

Neighbors of tag "hp"	Neighbors of word "hp"
harry potter	hewlett packard
hp movies	hp.com
hp books	hp computers
hp book quotes	hp company
harry potter facts	dell computers
hogwarts	hp printers

5.3 Forming interest segments

The goal of interest prediction is to identify groups of users with interest in certain topics. In this section we describe method for predicting user interests used in this study.

After we obtained categorized tags and keywords as described in the previous sections, the interest score for user u_i at time t in the k-th interest category was calculated as

$$u^t_{i,cat=k} = \sum_{feat \in \mathcal{A}_i} \alpha^{(t-t_{feat})}\ w_{feat}\ I(feat \text{ is of class } k), \tag{5.5}$$

where $\mathcal{A}_i$ is the set of extracted actions or keywords of user u_i contained in user's profile as described in Section 4.2, w_{feat} is a value of action or keyword feature (with 0/1 values for actions and counts for keywords), while the indicator function $I(\cdot)$ returns 1 if an action or a keyword extracted from a user activity is categorized into class k, and 0 otherwise (note that we postpone description of action categorization to Section 5.3.1). In addition, we used timestamp t_{feat}, representing day on which the activity happened, to exponentially decay less recent activities to account for passing interest (we used $\alpha = 0.99$ in our experiments).

Thus, the value of $u^t_{i,cat=k}$ represents an exponentially time-decayed count of all the activities in the k-th interest category. In order to effectively store user profiles for interest targeting and avoid storing all possible activities with their timestamps, we maintain a decayed sum for each category and update $u^t_{i,cat=k}$ daily. Using this approach we are able to qualify top K users in each category by sorting their interest scores $u^t_{i,cat=k}$, where the choice of K varies from campaign to campaign depending on advertiser's goals. Sev-

[11] http://youtu.be/ygn5oUBydfM, accessed June 2015

Table 5: Examples of interest inference based on categorized user features

User	Inferred interest	User profile
user 1	Arts & Entertainment/Movies	tag:{spoilers:30, shrek:18, hercules:12, cinderella:3, hobbit:123, hulk:21, pokemon:7, thor:58, ... disney:500, tarzan:8, marvel:385, wolverine:21, twilight:2, pixar:87, godzilla:1, x-men:53, ... pocahontas:4, avengers:134} txt:{aladdin:28, batman:10, bambi:12, movies:100} desc:{oscar:1, animation:12, comedy:1, movie:1, dvd:1}
user 2	Style & Fashion	tag:{ womensfashion:110, curls:6, fashiondiaries:133, redhair:2, menswear:125, chanel:4 ... springfashion:50, style:132, streetstyle:132, hairstylist:134, dapper:3, mensfashion:124} txt:{fashion:108}
user 3	Food & Drink	tag:{food:11, dessert:4 , soup:1, brunch:1, fruit:2, chicken:3, smoothie:1, cake:2, breakfast:2, ... ginger:2, salad:5, avocado:1} txt:{food:16, meals:6} follows:{user4542:1, user84852:1, user9332:1, user4524424:1 }
user 4	Home & Garden	tag:{daisies:2, kitchen:20, chair:3, art:81, outdoor:20, chandelier:12, lamp:8, window:2, bath:1 ... floral:17, home:3, wildflowers:1, flowers:102, interior:201, tree:1, flower:49, table:1, stairs:2, ... bedroom:56, wood:2, bathroom:26} txt:{garden:32, interior:17, home:41}
user 5	Automotive/Motorcycles	tag:{cars:24, ride:9, vehicle:22, riding:8, road:18} txt:{bike:8, motorcycle:10, riding:5, ride:9, road:10, vehicle:18, bikes:6, bicycle:2, scooter:1}

Table 6: A/B test results on 10% of user population

Campaign	Control	Targeted
Home & Garden	–	+9.71%
Style & Fashion	–	+42.53%
Sports/Outdoor Sports	–	+19.86%
Arts & Enter./Television	–	+24.37%
Arts & Enter./Video Games	–	+19.02%
Pets/Dogs	–	+27.21%
Arts & Enter. (campaign no. 1)	–	+9.08%
Arts & Enter. (campaign no. 2)	–	+6.54%

eral examples of qualified user profiles are given in Table 5. Note that a user may be qualified into more than one interest category. When the system was deployed in the production, each user was assigned to 13 categories on average.

5.3.1 Leveraging the follower graph

In order to target Tumblr users who do not create much content, but actively follow and engage with other blogs, we leverage the follower graph to create additional categorized features. In particular, using equation (5.5) we identify users with high value of $u^t_{i,cat=k}$ (which we term *influencers*). Then, following and liking posts created by influencers in the k-th category can serve as additional evidence of one's interest in that category.

To implement this idea, we labeled 5% of users with the highest interest score in a certain category as influencers, and categorized "follow", and "like" actions directed towards such users into that category. Then, we recompute equation (5.5) with an extended set of categorized activities that includes the categorized actions. This effectively expands the interest segments with users that are not content producers, but mostly act as consumers of content.

5.4 Results

In order to evaluate the generated user interest segments, we performed online A/B testing and worked with several advertisers who ran concurrent interest-targeted and untargeted campaigns. We tracked user engagement with their ads in terms of sponsored post likes, reblogs, and follows, and present the results for 8 targeting campaigns in Table 6. We observed an average increase of 20% in user engagement with sponsored posts in comparison to untargeted campaigns (aggregated over 3 metrics), representing a significant improvement over the baseline approach.

6. GENDER PREDICTION

In this section, we explain the details of our gender prediction model, based on the user profiles described in the previous sections. We first describe the generation process of a golden set of labeled users, which is used to train a predictive model that generalizes well on the remaining unlabeled users. This is followed by a description of the classification model and a discussion of the empirical results.

6.1 Collecting ground-truth labels

In order to train machine learning method for gender prediction, in addition to user profiles we also require labels that present the ground truth (i.e., "male" or "female"). However, Tumblr does not collect gender information during sign-up, leaving open the question of how to obtain such data.

To address this problem, we propose to leverage highly informative blog description data in order to infer user gender. In particular, users often declare their names in their blog descriptions, as illustrated in Figure 1. To extract the declared names, we used several regular expression rules that we found to result in high precision. The obtained results from a large set of name-matching regular expressions were editorially tested for quality. It was found that regular expressions reported in Table 7 yielded the most reliable extracted names (valid names were extracted in more than 95% of the cases).

Next, in order to generate the ground truth, we used US census data of popular baby names[12] from year 1880 to 2013 to create "name → gender" mapping. More specifically, we used male/female empirical ratios as soft labels, with 1 indicating 100% confidence in male and 0 indicating 100% confidence in female name. This approach resulted in 564 thousand female and 395 thousand male users found.

6.2 Proposed approach

Let $\mathcal{D}_g = \{(\mathbf{x}_i, y_i), i = 1, ..., N\}$ denote our gender data set, where N is the total number of labeled users, $\mathbf{x}_i$ is a K-dimensional user feature vector, and $y_i \in [0, 1]$ is a soft gender label. The feature vectors were generated from the

[12] www.ssa.gov/oact/babynames/limits.html, 06/15

Table 7: Matching names in blog description

Pattern	Count
my name is *	783,564
my name's *	291,811
me llamo *	47,663
the name's *	38,065
mi nombre es *	9,751
mi chiamo *	9,181
mein name ist *	1,025
meu nome e *	512
mon nom est *	215
mio nome e *	185

Table 8: Accuracy of gender model on hold-out set

Gender	Precision	Recall
female	0.806	0.838
male	0.794	0.689

user profiles as described in Section 5.3 by setting $\alpha = 1$, which turns off the time-decay of feature counts (due to the fact that, unlike interest, gender does not fluctuate). To handle large feature counts, we normalized the values by applying log transformation: assuming that the count is x, we replaced feature value with $\log(1 + x)$.

Our goal is to learn a gender predictor, $f : \mathbf{x} \to y$. As a classification model, we used logistic regression, parameterized by weight vector $\mathbf{w}$. We assume that the posterior gender probabilities can be estimated as a linear function of input $\mathbf{x}$, passed through a sigmoidal function,

$$\mathbb{P}(y = 1|\mathbf{x}) = f(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{x}^{\mathrm{T}}\mathbf{w})}, \qquad (6.1)$$

and $\mathbb{P}(y = 0|\mathbf{x}) = 1 - \mathbb{P}(y = 1|\mathbf{x})$. To estimate the parameters $\mathbf{w}$, we minimize the following loss function,

$$\min_{\mathbf{w}\in\mathbb{R}^K} \frac{1}{N}\sum_{i=1}^{N} \big(y_i - f(\mathbf{x}_i, \mathbf{w})\big)^2 + \lambda\|\mathbf{w}\|_1, \qquad (6.2)$$

where hyper-parameter λ controls the ℓ_1-regularization, introduced to induce sparsity in the parameter vector and reduce the feature space to a subset of features that are most predictive. In addition, we experimentally observed that the model generalizes better when we trained an initial model with ℓ_1-regularization to find which features have non-zero weight, and then do another round of training without ℓ_1-regularization, by only using features with non-zero weights from the first round to learn a better classifier.

Given a trained LR model, the posterior class probabilities are estimated as $f(\mathbf{x}_i, \mathbf{w}) \in [0, 1]$. Then, the predictions are made by thresholding, as $\hat{y}_i = \mathrm{sign}(f(\mathbf{x}_i, \mathbf{w}) - \theta)$, where threshold $\theta \in (0, 1)$ is set to ensure desired precision and recall according to advertiser's specific requirements.

6.3 Results

To evaluate accuracy of our gender prediction framework, we trained a logistic regression model on 70% of the golden set and tested on the remaining 30%. We used Vowpal Wabbit [13] implementation on Hadoop to train the model. To illustrate the performance of our gender classifier, the performance results in terms of precision and recall measures are presented in Table 8. The threshold value θ was set to a value which ensured precision of 0.8.

Table 9: Editorial evaluation of random user predictions

Prediction	# correct	# wrong	# not sure
female	429	4	298
male	144	5	127

In addition to evaluation on the hold-out set, we editorially evaluated gender predictions on the unlabeled data set of user profiles. We randomly picked 1,007 gender predictions from the population of 64.1 million users and asked editors to visit their profiles and verify their gender. They were instructed to mark our predictions as "correct", "incorrect", or "not sure". The "not sure" grade is to be used when the visual inspection of a profile is inconclusive, as we found was often the case. The editorial judgment came back with 573 "correct" (429 females and 144 males), 9 "incorrect", and 425 "not sure" grades (see Table 9). The fact that there are so many "not sure" grades indicates that in many cases it is hard to infer the gender even after manual efforts, further indicating the benefits of the proposed approach and its superior performance in comparison to humans. Finally, we retrained the model with 100% of the golden set and deployed it in Yahoo production systems. A demonstration video of gender predictive tags is available online[13].

7. DEPLOYED SYSTEM

To keep up with large number of daily activities, we implemented daily scoring of users on Yahoo production servers. We store the activity raw counts as well as decayed counts in Hive tables[14] for efficient retrieval. The decayed counts used in interest prediction are updated on a daily basis by multiplying the old feature values by the decay factor α and adding new activities. In order to infer the gender of new users we implemented daily scoring by leveraging MapReduce on Hadoop[15]. Both interest and gender models are retrained on a regular basis.

After thorough editorial evaluation of the inferred gender and interest targeting, both targeting frameworks were enabled through Gemini self-serve tool. Advertisers can choose to use gender and/or interest targeting with custom segment sizes, allowing for effective targeting campaigns.

8. CONCLUSION

We presented the steps in the development of a large-scale Tumblr gender and interest targeting framework, where we used historical Tumblr activities to create rich user profiles. We described the methodology, including a recently proposed semi-supervised neural language model, as well as the high-level implementation details behind the deployed system. Currently, our gender and interest predictions cover users that generate more than 90% of overall daily activities on Tumblr, and are heavily leveraged by advertisers. In our ongoing work, we are concentrating on creating custom keyword-targeted advertising segments specifically tailored for a particular advertiser, which includes work on addressing the problems of keyword discovery and expansion.

[13] https://youtu.be/jXGJ0TpOlhg, accessed June 2015
[14] https://hive.apache.org, accessed June 2015
[15] https://hadoop.apache.org, accessed June 2015

9. REFERENCES

[1] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 114–122, 2011.

[2] J. Alba, J. Lynch, B. Weitz, C. Janiszewski, R. Lutz, A. Sawyer, and S. Wood. Interactive home shopping: consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *The Journal of Marketing*, pages 38–53, 1997.

[3] N. Barbieri, F. Bonchi, and G. Manco. Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1266–1275. ACM, 2014.

[4] F. Bonchi, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. Efficient query recommendations in the long tail via center-piece subgraphs. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 345–354. ACM, 2012.

[5] A. Z. Broder. Computational advertising and recommender systems. In *Proceedings of the ACM conference on Recommender systems*, pages 1–2. ACM, 2008.

[6] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu. What is tumblr: A statistical overview and comparison. *ACM SIGKDD Explorations Newsletter*, 16(1):21–29, 2014.

[7] R. K. Chellappa and R. G. Sin. Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management*, 6(2-3):181–202, 2005.

[8] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In *WWW*, pages 271–280. ACM, 2007.

[9] N. Djuric, M. Grbovic, V. Radosavljevic, N. Bhamidipati, and S. Vucetic. Non-linear label ranking for large-scale prediction of long-term user interests. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[10] N. Djuric, V. Radosavljevic, M. Grbovic, and N. Bhamidipati. Hidden conditional random fields with distributed user embeddings for ad targeting. In *IEEE International Conference on Data Mining*, 2014.

[11] D. Essex. Matchmaker, matchmaker. *Communications of the ACM*, 52(5):16–17, 2009.

[12] M. Grbovic, N. Djuric, V. Radosavljevic, N. Bhamidipati, J. Hawker, and C. Johnson. querycategorizr: A large-scale semi-supervised system for categorization of web search queries. In *International World Wide Web Conference (WWW)*, 2015.

[13] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *The Journal of Machine Learning Research*, 10:777–801, 2009.

[14] A. Majumder and N. Shrivastava. Know your personalization: Learning topic level personalization in online services. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 873–884, 2013.

[15] U. Manber, A. Patel, and J. Robison. Experience with personalization on Yahoo! *Communications of the ACM*, 43(8):35, 2000.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[17] D. Riecken. Personalized views of personalization. *Communications of the ACM*, 43(8):27–28, 2000.

[18] D. Shin, S. Cetintas, and K.-C. Lee. Recommending tumblr blogs to follow with inductive matrix completion. In *RecSys 14 Poster Proceedings*, 2014.

[19] T. Singh, L. Veron-Jackson, and J. Cullinane. Blogging: A new play in your marketing game plan. *Business Horizons*, 51(4):281–292, 2008.

[20] S. K. Tyler, S. Pandey, E. Gabrilovich, and V. Josifovski. Retrieval models for audience selection in display advertising. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 593–598. ACM, 2011.

[21] C. Yi, T. Lei, I. Yoshiyuki, and L. Yan. What is tumblr: A statistical overview and comparison. arXiv:1403.5206v2, 2014.