User Profiling on Tumblr through Blog Posts

Kriti Pandey*, Ayush Mittal*
Indraprastha Institute of Information Technology, Delhi (IIIT-D)
Delhi, India
{kriti12050, ayush12030}@iiitd.ac.in

***Abstract*— Tumblr, a microblogging service, has garnered a lot of prominence in recent past with 269.3 million blogs and 125.9 billion posts which makes it a hub for creative content. Despite this drastic growth of Tumblr since its launch, not a lot of scholarly work has been done on it. This work is an attempt to comprehensively profile the blogger in terms of age, gender and relationship status on Tumblr. The analysis conducted in this paper gives an approach for interest categorization of blogs based on topical analysis and writing characteristics. A normalized weighted representation of the most talked about categories for a blog was inferred based on the correlation between topical words and defined category related terms. The trained models predict age, gender and relationship status of the user with an accuracy of 82.37%, 79.61% and 76.52% respectively. To the best of our knowledge, this is the first attempt to profile users on Tumblr.**

Keywords—social networks, Tumblr, attribute prediction, user profiling

I. INTRODUCTION

Online social network platforms have now become a very popular channel for people to express and exchange their thoughts, interests and activities with their social connections. Facebook (1.55 billion monthly active users)[1] [1], Twitter (320 million monthly active users)[1] [2], LinkedIn (400+ million registered users)[1] [3] are some of the most popular social networks for connecting with people and establishing contacts. With an upsurge in number of people engaging in social media activities, these platforms have refurbished their basic structure and gone beyond just sharing and posting text.

A Web service like Flickr became immensely popular as an image and video hosting website, for the users to share and embed personal photographs. Pinterest is another image based social network wherein users can save, upload, manage or sort images known as pins according to their interest. Although fairly new to the social media clan, Instagram is also gaining a lot of traction with over 80 million photo uploads per day and 400+ million monthly active users [4]. Recently, a lot of people have started using these platforms to emulate their real life relations and establish a social presence.

Other than these text-intensive and image based social networks, there are some notable blog sharing networks too like Blogger, LiveJournal etc. It has been shown in [5] that blogging is an innate expression of one's true self and shows a lot about the attitude and conviction of the writer. Other than the traditional blogging community, microblogging services also exist wherein the content is typically smaller in aggregation. Twitter, a very popular microblogging platform poses a limit of 140 characters for each tweet. Very similar to Twitter, Tumblr is also one of the most widely used microblogging network and has gained a lot of vogue in recent years. It has been listed as the fastest growing social platform overtaking Instagram [6] and is the second largest microblogging service after Twitter [7]. Unlike Twitter, Tumblr doesn't have any limit on the length of the post and also allows a user to post multimedia content like photo, text, audio, video, quote, chat, question and link. Due to all the above stated reasons, Tumblr is a potential hub for a lot of creative and diversified content and this sets it apart from its alike. While signing up on Facebook requires a user to fill in first name, last name, e-mail address or mobile number, date of birth and gender, it only takes an e-mail address and a unique username to register for a blog on Tumblr. Any user on the graph network of Tumblr is represented by just a blog name. It is completely at the user's discretion to furnish any personal details on the blog. With no limitation on post length and no other plausible identification, users are very candid while posting their blogs and choose to express with no restrain.

This paper is an attempt to study whether the wealth of rich and high quality content available on Tumblr network can be correlated to generate a user profile in terms of age, gender and relationship status of the blogger. In simpler words we intend to infer these attributes for a user from the blog posts he or she has written. Since users on Tumblr choose whether or not to reveal details about themselves, any ability of this kind to infer such characteristics of these users would certainly help in content recommendation, scoped sharing or sponsored advertisements. However, this has a greater privacy implication wherein a blogger's anonymity may just be an illusion and is regardless of what he or she chooses to reveal. The major objective of this work is to analyze if expression of thoughts and personal communication in form of seemingly harmless blog posts when combined with semantic techniques can lead to giving away of some personal information.

To get a deeper understanding on this subject matter a total of 329,526 Tumblr blogs were collected. Out of these, 17,073 blogs with self-reported attributes in their profiles were extracted. Further, a total of 24 broad categories and their related terms were modelled for topical analysis. Each blog with self-reported attributes was mapped to one or more of these categories post applying Latent Dirichlet allocation

*Authors contributed equally to this work.

[1]The statistics are as published by the social networking sites in September, 2015

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

(LDA) algorithm [8] on all the text posts pertaining to that blog. Having established the ground truth for interest categories and attribute mapping, latent attributes like age, gender and relationship status were predicted using multiple classification techniques. Different classifiers perform differently based on the feature sets used for training.

The rest of the paper is organized as follows: the related work has been discussed in Section 2. The data collection process is given in Section 3 followed by the analysis methodology discussed in Section 4. Corresponding results are covered in Section 5 and the paper is concluded in Section 6.

II. RELATED WORK

A. Inferring Atrributes on other social networks

There has been a considerable amount of work done in inferring latent attributes or author profiling on different social media platforms using various techniques. Private attributes of users have been inferred based on their music interests in [9] particularly for Facebook. The list feature and its metadata on Twitter has been used to obtain characteristics of a twitter handle in [10]. Latent attributes have been inferred for a larger section of users on social media using declared attributes of only a fraction of the users in [11]. In [12] gender, age, and political affiliation of users on Twitter have been predicted using the principle of Homophily which states that a user can be characterized using his or her neighborhood attributes since people in virtual space look out for others with shared or similar interests. Several other attempts have also been made to infer private attributes on social media platforms mainly using the conventional Naïve Bayes classifiers [13-15].

Considering the numerous studies done in inferring private information about users on various other social networks, not much of similar work has been extended to Tumblr. One can say that the available unstructured data and dearth of concrete information about users is a major challenge in reproducing the above discussed works on Tumblr. This work in itself is a novel attempt to profile users on Tumblr based on augmented interest categories and writing characteristics.

B. Tumblr Research

Considering the phenomenal growth of Tumblr as a social network since its inception, not a lot of work has been done in this field. A large scale statistical analysis on Tumblr encompassing its social structure, user generated content analysis and characterizing the reblogging pattern and behavior has been done in [7]. An inductive matrix based blog recommendation method using multiple rich and varied sources of evidences from social media, user blogging activity, etc. has been presented in [16]. A close analysis of Tumblr as a channel for information diffusion and investigation of structural characteristics of cascades has been carried out in [17]. Reference [18] studies the motivations of fandom users to choose Tumblr over other social media platforms and investigates the way these users interact within Tumblr community.

In a closely related work [19], a gender prediction framework on Tumblr has been proposed by training a logistic regression model in contrast to our work wherein an entire user profile has been predicted comprising of gender, age group and relationship status.

III. DATA COLLECTION

In this section a brief description of the methodology deployed in data collection for this work has been given along with a characterization of the datasets used. The publicly available API for Tumblr has been used to collect each of the dataset as described in the subsections presented below. The entire data collection process spanned over a period of 38 days from September 5, 2015 to October 12, 2015. The data collection was done using Amazon M4 EC2 instance with 32 Gigabytes of RAM and 8 vCPUs. The entire data collection process has been well represented in Fig. 1.

A. Blogs, Blog Info and Blog Posts Dataset

A breadth first search (BFS) crawler was programmed in Python for initiating the data collection of blogs which was fed with 5 viral Tumblr blogs [20] as seeds. The crawler first collected the publicly exposed likes of these seeds and then repeatedly extracted the "likes of likes". In this process a total of 329,526 blogs were collected which we'll call as the Blogs dataset. Alongside, the meta information, which included the title and description of all these blogs, was also gathered and stored as the Blog Info dataset.

Next, the Blogs dataset was crawled and all the public posts corresponding to each blog were retrieved. The title, body and tags for all the text posts were stored in the Blog Posts dataset. All the other types of posts were aggregated in another dataset with its corresponding type. Fig. 2 shows the distribution of post types based on a total of 422,382,716 posts collected. Clearly, the photo content dominates the distribution which is in accordance with the previous works done on Tumblr [7, 19].

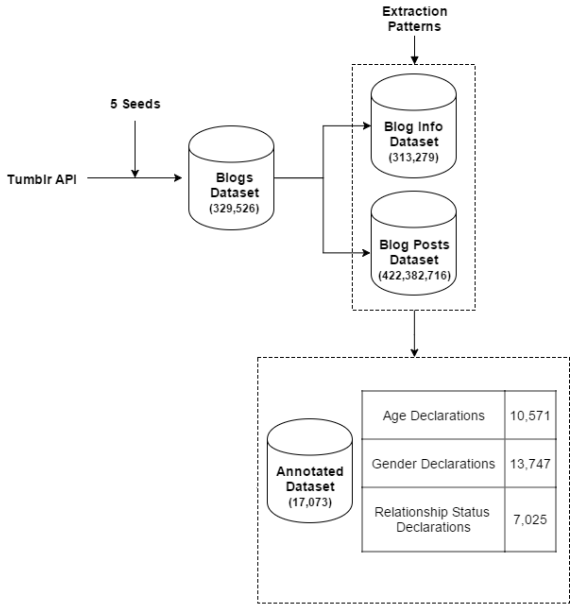

Fig. 1. Procedural depiction of the data collection process. The table alongside the annotated dataset shows the characterization for all the annotations available across attributes.

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

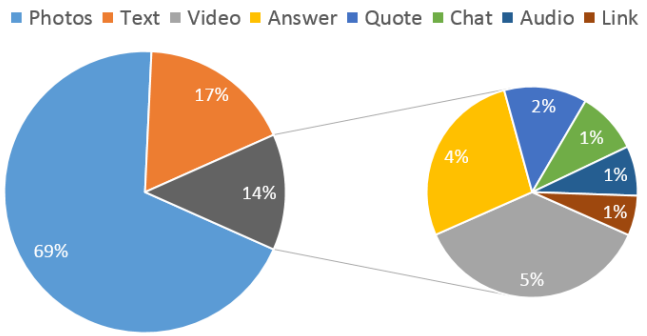

Fig. 2. Distribution of Tumblr posts based on type

B. *Annotated Dataset*

Many a times users explicitly declare their name or other personal information in the blog distribution or their posts which can help in uniquely characterizing a user. For example- "celebrated my 20th birthday", "I am in love with a guy", "in a relationship", etc. We have leveraged this fact and applied multiple extraction patterns on the Blog Info and Blog Posts datasets to retrieve a set of those blogs which have self-reported such attributes in their profiles. Some of the extraction patterns used are given in Table I. This has been named as the Annotated dataset wherein each blog is mapped to its stated demographics. In this process a total of 17,073 blogs were collected with explicit mentions of their demographics of one or more attributes. In each of the datasets discussed in subsection A of Section 3, only the information relevant to annotated blogs was kept. Post discarding, the Blog Posts dataset consists of 3,136,753 text posts which will be used for further analysis.

IV. METHODOLOGY

A. *Data Processing*

The Blog Posts dataset consists of the title, post and tags of a blog post which are returned in HTML format. For creating a meaningful depiction of these posts, we stripped off the HTML tags and constructed its plain text representation followed by elimination of common stop words. Further, only those posts were considered for analysis whose at least 50% of the words occured in a standard English dictionary. These posts were then passed to a stemmer to obtain the root words. For example eat, eats, eating, eaten are stemmed to the root word eat.

TABLE I. EXTRACTION PATTERNS USED FOR AGGREGATING ANNOTATED DATASET

Attribute	Declaration Type	Example
Age	explicit declaration	"I'm 25 years old"
	plain number to depict age	"I'm 20" ," life in 20's is great"
	declaring year of birth	"Born in 1990"
	ambiguous declaration of range	"I may be an adult"
Gender	explicit declaration	"I am a girl"
	declaration of relation to another	"I am a mother of two kids"
	contextual declaration	"boy's night out", "girl's sleepover"
Relationship Status	explicit declaration	"in relationship", "single", "married"
	contextual declaration	"my boyfriend is out of town", "missing my wife", "being single sucks"

B. *Topic Modeling and User Categorization*

For topical distribution of blog posts of the annotated users, two standard taxonomy hierarchies- Open Directory Project [21] and AlchemyAPI [22] were considered. A total of 24 categories were finalized based on the top common categories given by two taxonomies and the related words were taken from their sub category levels. These categories with their related words are listed in Table II.

This topical distribution considers the fact that different types of people have different augmented interests. This work tries to model these interest differences to predict age, gender and relationship status of a user in labels as shown in Table III.

TABLE II. CATEGORIES AND SOME CATEGORICAL WORDS

Category	Keywords
Art and Crafts	art, embroidery, folding, needle work
Automotive	car, bike, sedan, hatchback
Career	entrepreneurship, startup, engineer, job
Education	school, college, homework
Entertainment	hollywood, music, dance, thriller, tv, comedy
Environment and Nature	weather, climate, tree, animal, global warming
Family	pregnancy, motherhood, babycare
Fandom	anime, manga, marvel, comic
Fashion	model, glamour, costume
Finance	fund, insurance, credit card, investment
Food/Drink	appetizer, cocktail, cuisine, beer, dessert
Gaming	nintendo, xbox, dota, fifa, EA sports
Health	allergy, fitness, medicine, fever, gym
Hobbies	shopping, photography, stamp, coin
Literature	best seller, poetry, book, magazine
Personal Moods/ Personal Life	relationship, depression, love, friendship, party
Politics	election, political party, law
Religion	god, christianity, catholic, hindu, muslim, atheism
Science and Technology	gadget, smartphone, laptop, windows
Sexual Content	slut, porn, sex, nude
Social Media	selfie, hashtag, instagram, instameet, filter
Societal Issues	terrorism, LGBT, corruption
Sports	football, cricket, baseball, knicks
Travel	honeymoon, tourist, camping, hotel

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

TABLE III. POSSIBLE CLASS LABELS FOR EACH ATTRIBUTE

Class	Possible Class Labels
Age	16-20 years
	21-25 years
	26-30 years
	30+ years
Gender	Male
	Female
Relationship Status	Single
	Committed

As already stated, there is no limit on the length of posts on Tumblr and hence it generates a lot of high quality content [7]. This enables us to topically distribute these posts to get meaningful representation. LDA algorithm with number of topics as 7 was applied[2] on all posts of each blog in the Annotated dataset to see the topic of interest of the users. We took top k-topics for each user and mapped them to one or more of the 24 categories by correlating the topical terms of each user and the related words of the categories. As per the constructed mapping, normalized weights were given to each category. A word cloud of most frequent words used in a blog has been given as an example in Fig. 3 for interest depiction.

C. *Attributes Prediction and Feature Set*

Having established the procedure to infer topic of interests for each user, three classification techniques were used to train multiple models, one each, for predicting age, gender and relationship status of the user. The classifiers were trained based on augmented interest categories (F1) as inferred from the user's blog posts and writing characteristics (F2) as shown in Table IV. The classification tasks were performed considering the fact that the information furnished by 17,073 blogs is self-declared and hence in most cases would be true.

TABLE IV. FEATURES CLASSIFIED INTO INTEREST BASED AND WRITING CHARACTERISTIC.

Feature Type	Feature	Explanation
Interest Based Feature (F1)	Topical Interest	Normalized Weighted representation of top interests as inferred from the blogs. Example: 0.8* Food/Drink + 0.1*Gaming + 0.1*Finance
Writing Characteristic Feature (F2)	No of slangs	Number of words/phrases not in english dictionary, used in an informal way. Example: wassup, sup, gotchya, cya
	No of emoticons	Number of standard emoticons used. Example: :), ;), <3
	No of acronyms	Number of internet acronyms used. Example: LOL, ROFL, ASAP

[2]The number was decided empirically by repeated experiments

Fig. 3. Word cloud of most frequent words used in posts of a user mapped to category Food/Drink, Science and Technology, Gaming and Travel.

V. RESULTS

The three classifiers applied on the annotated dataset were Naive Bayes, J48 Decision Tree and Random Forest with 10-fold cross-validation. It is assumed that since all observations are used for both training and validation exactly once, it would reduce the training error due to sparse class sizes. An average accuracy of 80.29% for age, 76.84% for gender and 71.93% for relationship status was obtained when trained on both feature sets (F1+F2). The models when trained on both features perform better than when trained only on F1 or F2 as is evident from the results summarized in Table V. However, relationship status shows the least increase in accuracy across all classifiers as we move from training it on F1 to both F1 and F2. Hence it can be said F2 is least informative for this attribute. In each of the subsection below, we have presented one classifier based on the maximum accuracy obtained for that attribute on the validation set.

A. *Age Prediction*

The training dataset comprised of 1,487 users in the age group 16-20 years, 5,439 users in the age group 21-25 years, 2,637 users in the age group 26-30 years and 1,008 users aged 30+. The maximum accuracy of 82.37% was achieved by Random Forest using both F1 and F2.

TABLE V. CLASSIFICATION RESULTS FOR NAIVE BAYES, J48 DECISION TREE AND RANDOM FOREST (NUM OF TREES=100)

Model	Feature Set	Accuracy (in %)		
		Age	*Gender*	*Relationship Status*
NB	F1	74.91	70.57	65.67
	F2	62.39	56.92	47.89
	F1 + F2	78.29	73.38	66.08
J48 DT	F1	76.57	72.15	72.78
	F2	64.33	63.57	55.26
	F1 + F2	80.23	**79.61**	73.19
RF	F1	80.77	75.81	76.06
	F2	68.52	64.59	61.08
	F1 + F2	**82.37**	77.54	**76.52**

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

B. Gender Prediction

The training dataset comprised of 5716 male users and 8031 female users. The maximum accuracy of 79.61% was achieved by Decision Tree using both F1 and F2.

C. Relationship status Prediction

The training dataset comprised of 3289 single users and 3736 users in relationship. The maximum accuracy of 76.52% was achieved by Random Forest using both F1 and F2.

VI. CONCLUSION

This work is a maiden attempt to profile Tumblr bloggers in terms of latent attributes like age, gender and relationship status. Topical analysis of blog posts of the user helps in interest categorization of the user. Based on these augmented interest categories and other writing characteristics, multiple classification techniques yield varied accuracy results. The highest accuracy for age is given by Random Forest, for gender is given by J48 Decision Tree and for relationship status is given by Random forest, where all the classifiers are trained on F1 and F2 together. However feature set F2 is least informative for relationship status as the increase in accuracy for all classifiers for this attribute, when trained only on F1 and when trained on both F1 and F2 together, is not significant and lowest amongst all other attributes.

REFERENCES

[1] http://investor.fb.com/releasedetail.cfm?ReleaseID=940609

[2] https://about.twitter.com/company

[3] https://press.linkedin.com/about-linkedin

[4] https://www.instagram.com/press/

[5] T. Mortensen and J.Walker. Blogging thoughts: Personal publication as an online research tool.Researching ICTs in Context, A. Morrison, Ed. InterMedia Report, Oslo, Norway, 2002

[6] http://techcrunch.com/2014/11/25/tumblr-overtakes-instagram-as-fastest-growing-social-platform-snapchat-is-the-fastest-growing-app/

[7] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu, "What is Tumblr: A Statistical Overview and Comparison," ACM SIGKDD Explorations Newsletter - Special issue on big data, vol. 16, pp. 21-29, June 2014. doi 10.1145/2674026.2674030

[8] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993- 1022, January 2003.

[9] A. Chaabane, G. Acs, and M. Kaafar, "You Are What You Like! Information Leakage Through Users' Interests," In Proc. Annual Network and Distributed System Security Symposium, 2012.

[10] N.K. Sharma, S.Ghosh, F. Benevenuto, N. Ganguly, and K.Gummadi, "Inferring Who-is-Who in the Twitter Social Network," In Proc. ACM Workshop on online social networks(WOSN'12), pp. 55-60, 2012. doi 10.1145/2342549.2342563

[11] A. Mislove, B. Viswanath, K.P. Gummadi, and P. Druschel, "You Are Who You Know: Inferring User Profiles in Online Social Networks," In Proc. ACM international conference on Web search and data mining(WSDM'10), pp. 251-260, 2010. doi 10.1145/1718487.1718519

[12] F.A. Zamal, W. Liu, and D. Ruths, "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors," In Proc. AAAI Conference on Weblogs and Social Media(ICWSM 2012), June 2012.

[13] J. He, W. W. Chu, and Z. V. Liu, "Inferring privacy information from social networks," In Proc. IEEE international conference on Intelligence and Security Informatics, pp. 154-165, 2006. doi 10.1007/11760146_14

[14] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Preventing private information inference attacks on social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 1849 - 1862, June 2012. doi 10.1109/TKDE.2012.120

[15] W. Xu, X. Zhou, and L. Li, "Inferring privacy information via social relations," IEEE Conference on Data Engineering Workshop (ICDEW 2008), pp. 525-530, April 2008. doi 10.1109/ICDEW.2008.4498373

[16] D. Shin, S. Cetintas, and K.C. Lee, "Recommending Tumblr Blogs to Follow with Inductive Matrix Completion," In Proc. ACM Recommender Systems conference (RecSys 14), Poster Proceedings, 2014.

[17] N. Alrajebah, "Investigating the Structural Characteristics of Cascades on Tumblr," In Proc. IEEE/ACM Conference on Advances in Social Networks Analysis and Mining 2015(ASONAM '15), pp. 910-917, 2015. doi 10.1145/2808797.2808814

[18] S. Hillman, J. Procyk, and C. Neustaedter, "alksjdf;lksfd': Tumblr and the Fandom User Experience," In Proc. Designing interactive systems (DIS '14), pp. 775-784, 2014. doi 10.1145/2598510.2600887

[19] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, and A Nagarajan, "Gender and Interest Targeting for Sponsored Post Advertising at Tumblr," In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), pp. 1819-1828, 2015. doi 10.1145/2783258.2788616

[20] http://mashable.com/2014/12/03/top-25-viral-tumblrs-2014/#Tu68PVtzCZqS

[21] www.dmoz.org

[22] www.alchemyapi.com/api/taxonomy/

978-1-5090-0082-1/16/$31.00 ©2016 IEEE