7-2014

On predicting religion labels in microblogging networks

Minh Thap NGUYEN
Singapore Management University, mtnguyen.2012@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2600428.2609547

On Predicting Religion Labels in Microblogging Networks

Minh-Thap Nguyen
Living Analytics Research Centre
Singapore Management University
mtnguyen.2012@smu.edu.sg

Ee-Peng Lim
Living Analytics Research Centre
Singapore Management University
eplim@smu.edu.sg

ABSTRACT

Religious belief plays an important role in how people behave, influencing how they form preferences, interpret events around them, and develop relationships with others. Traditionally, the religion labels of user population are obtained by conducting a large scale census study. Such an approach is both high cost and time consuming. In this paper, we study the problem of predicting users' religion labels using their microblogging data. We formulate religion label prediction as a classification task, and identify content, structure and aggregate features considering their self and social variants for representing a user. We introduce the notion of *representative user* to identify users who are important in the religious user community. We further define features using representative users. We show that SVM classifiers using our proposed features can accurately assign Christian and Muslim labels to a set of Twitter users with known religion labels.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Religion Prediction; User Profiling; Social Networks

1. INTRODUCTION

Motivation. In many consumer and social applications, user attributes are required to suggest relevant and interesting products, content, services and social links. Among the many user attributes is religion which has proven to be very important in determining how users behave. Religion influences how users form preferences, interpret events around them, and develop relationships with others. In the past, religion labels are obtained by large scale user surveys run by government agencies or large businesses with or without financial incentives (e.g., lucky draws, direct discounts, etc.). These survey efforts are generally effective but are intrusive and time consuming. On the other hand, online social media users have their religion attributes embedded in the content and interaction with other users. It is therefore interesting to recover these users' religion labels from their social media data, and to do so accurately.

Objectives. In this paper, we attempt to predict users' religions using their microblogging data. This task has not been addressed so far and datasets with religion labels are not publicly available. The task is particularly interesting for a user community that has a mixture of users with different religious beliefs. Our research begins with gathering a dataset covering a set of 111,767 Twitter users located in Singapore, their follow relationships, and their tweets. This Twitter dataset allows us to explore both content and structure features relevant to users' religious beliefs. This also distinguishes our user label prediction research from previous works which consider random sets of users, i.e., who are strangers to one another. We manually annotate the religion labels of over one thousand users who declare religion beliefs in their biographies. Using these labeled data, we are able to evaluate methods that predict user religion labels.

User religion prediction task is challenging. Most users do not reveal their religion labels explicitly. Out of the set of Singapore users we are able to identify only 1029 users, or $<$ 1%, who mention their religions clearly in their biographies. Vast majority of them do not. The sparsity of labeled data is even more severe for religions that have very few believers. Label sparsity poses several challenges to the prediction problem. Firstly, there are few labeled users for training classifiers. Secondly, even with labeled users, we may also have insufficient content and interaction data generated by some of them to learn an accurate classifier.

Related Works. Despite no previous works on predicting user religion, there are other related works on mining online social media user attributes including political affiliation, gender, ethnicity, and country. Bhargava and Kondrak performed language classification of people names using word and n-gram name features [3]. Pennacchiotti and Popescu proposed a classification method combining gradient boosted decision tree and graph updating to perform classification of user political affiliations, ethnicity and favorite businesses [6]. Their decision tree classifiers represent each user by their profile, tweeting behavior, linguistic content and social network features. It was then observed that some user attributes are harder to classify than others. Al Zamal et al. [1] showed that using neighbors' features only to predict age and political affiliation outperform using user

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609547 .

Table 1: Singapore Twitter Dataset

# total users	111,767
# total follow links	1,770,272
# labeled as Christians	581 (53.9%)
# labeled as Muslim	448 (41.6%)
# total labeled users	1029 (100%)

features only. This can be attributed to attributes' high assortativity. Rao et al. [7] proposed two sets of features to predict user gender, age, region origin and political affiliation. The first set of features are socio-linguistic words. The second set of features are unigrams and bigrams of the tweet text. Our work differs from the above works in several ways. First, we focus on user religion prediction which has not been studied earlier. Secondly, we approach the prediction task for a given user community as opposed to a random set of users who represent only sub-clusters of a larger community.

Contributions. We summarize our contributions as follows:

- We construct a very large user community consisting of more than 111K users that belong to a community, and assign the religion labels of about one thousand users so as to study the user religion prediction problem. This is also the first time the task is studied for a large user community.
- We systematically extract different types of user features covering both content and structure aspects of Twitter data. The content word features are specially selected to be relevant to the religion class labels. The structure features are derived from the follow relationships among users. We propose a novel *representative user measure* that allows us to determine users important among users sharing the same religion. Based on this measure, we derive content and structure features that improve the prediction accuracies.
- Using our proposed user features, we are able to train highly accurate SVM classifiers that yields F1-scores larger than 0.8 for both the Christian and Muslim labels. Such an accuracy level makes the classifiers useful in different real world applications.

2. DATASET AND RELIGION LABELS

Twitter user community. We crawled the Twitter data generated by about 110K users with Singapore specified as their profile location in July 2012. These users were identified by first constructing a set of well known Singapore user accounts as seeds. We then found other Singapore user accounts connected to the seeds. The process was repeated several times before reaching the above user count. The dataset consists of users' profiles, tweets, and follow links among the users. Table 1 shows the basic statistics of this dataset.

User religion labeling. There are five main religions in Singapore, namely, Buddhism, Taoism, Christianity, Islam, and Hinduism. When we attempted to assign religion labels to Singapore users, we found very few self-declared Buddhists, Taoists and Hindu's in our dataset. We therefore focus on labeling Christians and Muslims only. We manually searched religion specific keywords in the biography of users as shown in Table 2. This was carried out by first extracting users with biography fields containing any of the keywords, followed by manually judging the religion label for each extracted user based on their biographies. Only those who clearly state their religion were then assigned the Christian or Muslim label. As shown in Table 1, we managed to label 581 Christians and 448 Muslims. This composition structure is quite different from religion composition reported in the 2010 Singapore population census which shows Christians, Muslim, Buddhists and Free-thinkers represent 18.3%, 33.3%, 14.7% and 17% respectively. This suggests that Christian and Muslim users have higher propensity to share their religious beliefs online. Buddhist and free-thinker users in contrast are more reserved.

Table 2: Selected keywords for assigning religion labels

Religion	Selected Keywords
Christian	*jesus, christ, protestant, catholic, church*
Muslim	*allah, muslim, islam, mosque*

3. REPRESENTATIVE USERS

In a social network, some users may appear to be more important than others for a given religion. We are interested in determine these important users whose connections may help in the prediction of religion labels. The standard measures that characterize user importance in a network include degree centrality and pagerank. These measures however assume user importance is independent of their affiliation labels. For user religion prediction, we are interested in measures based on the user's importance in each affiliation group. We define three new measures, called *label-degree*, *label-indegree* and *representative indegree ratio*, as follows.

The **label-degree** of user u for label r, $deg(u, r)$, is defined as the number of neighbors (based on both in-links and out-links) of u from users with religion label r. The **label-indegree** of u for a religion label r, $\text{indeg}(u, r)$, is defined as the number of links to user u from users with religion label r. The **representative indegree ratio** of u, $\text{RInDeg}(u)$, is defined as: $\text{RInDeg}(u) = \frac{\text{indeg}(u, r_{max}(u))+\alpha}{\text{indeg}(u, r'_{max}(u))+\alpha}$ where $r_{max}(u) = \text{argmax}_r \text{indeg}(u, r)$, and $r'_{max}(u) = \text{argmax}_{r \neq r_{max}(u)} \text{indeg}(u, r)$. The parameter α is a smoothing constant whose purpose is to prevent undefined values when $\text{indeg}(u, r'_{max}(u)) = 0$. In the experiment, we set α to 1.

The representative degree ratio $\text{RInDeg}(u)$ (> 1) measures the dominance of a specific religion compared with other religions among u's followers. The larger $\text{RInDeg}(u)$, the more dominance is a specific religion among the followers. This suggests that u is very important among the followers with the specific religion affiliation. When u is followed by only users of a single religion, $\text{RInDeg}(u) = \frac{\text{indeg}(u, r_{max}(u))+\alpha}{\alpha}$. For $\alpha = 1$, $\text{RInDeg}(u) = \text{indeg}(u, r_{max}(u))+1$. When u is not followed by any users with religion label or when u is followed by equal largest groups of users with different religion labels, $\text{RInDeg}(u) = 1$.

Due to the use of already labeled users, the three measures are also quite different from the standard degree measure. A high degree user may not have high $\text{indeg}(u, r)$ for some r (and also $\text{RInDeg}(u)$) if the user is not followed by any labeled users.

Table 3: Top Representative Users (some names are truncated for space saving)

Rank	deg(u,r)		indeg(u,r)		RInDeg(u)	
	Christian	Muslim	Christian	Muslim	Christian	Muslim
1	konghee (141)	*TaufikB...* (126)	konghee (139)	*TaufikB...* (126)	konghee (140)	*Norfasarie* (59)
2	*STcom* (127)	*SoSinga...* (97)	*STcom* (127)	*SoSinga...* (97)	JosephPrince (91)	*MediaCo...* (32)
3	JosephPrince (90)	*banchot...* (93)	JosephPrince (90)	*banchot...* (93)	chcsg (64)	*iNasuha* (29)
4	*mrbrown* (89)	*sezairi* (69)	*mrbrown* (87)	*sezairi* (69)	nccsg (45)	*fadhilf...* (26)
5	*SoSinga...* (82)	WaktuSo... (69)	*SoSinga...* (82)	WaktuSo... (69)	joepurcell (36)	*didicazli* (25)
6	chcsg (63)	*STcom* (66)	chcsg (63)	*STcom* (66)	Celestfoo (35)	MuslimSG (24)
7	*Channel...* (52)	*FauzieLaily* (62)	*Channel...* (52)	*FauzieLaily* (62)	thezone... (33)	*HyrulAnuar* (24)
8	*Leticia...* (49)	*HadyMir...* (59)	*Leticia...* (49)	*HadyMir...* (59)	JianMingTan (27)	*DearAbdullah* (23)
9	nccsg (44)	*Norfasarie* (59)	nccsg (44)	*Norfasarie* (58)	garrettleejw (27)	*Fiza_O_* (23)
10	*Xiaxue* (44)	*Channel...* (43)	*Xiaxue* (44)	*Channel...* (43)	Chris_H... (25)	*TaufikB...* (21)

Table 4: Categorization of Features

	Content	**Structural**	**Aggregated**
Self	- binary terms - TF · IDF	- RInDeg - label degree for each religion r - label indeg for each religion r - label outdeg for each religion r - label Nindeg for each religion r - label Noutdeg for each religion r	- number of tweets - number/fraction of tweets containing selected terms for each religion r - $argmax_r$ indeg - $argmax_r$ outdeg - $argmax_r$ Nindeg - $argmax_r$ Noutdeg
Social	- binary terms from neighbors only - TF · IDF from neighbors only - binary terms from neighborhood - TF · IDF from neighborhood	- top k $\langle importance \rangle$ neighbors $\langle importance \rangle$ = {degree, RInDeg, indeg for each religion r }	- number/fraction of tweets from neighborhood containing selected terms for each religion r - avg/max/min of $\langle importance \rangle$ of top k $\langle importance \rangle$ neighbors $\langle importance \rangle$ = {degree, RInDeg, indeg for each religion r }

We apply the three proposed representativeness measures on all users in our Twitter dataset together with the set of labeled users shown in Table 1.

Table 3 lists the top ten users for each measure (with measure values given in parentheses). We can see that there are overlapping top representative users across religions by indeg and they include STcom, mrbrown, SoSinga..., Channel... which are owned by popular bloggers, and news agencies in Singapore (and are shown in italic fonts). RInDeg did a great job recognizing the top representative users unique to each religion, and these user accounts include: JosephPrince, konghee, chcsg, nccsg which belong to popular church leaders and churches, FauzieLaily, Norfasarie, MediaCo... are popular Muslim celebrities and news agencies, all based in Singapore. This observation indicates the promising benefit of using our proposed measure.

4. USER FEATURE REPRESENTATION

We represent each user using a vector of features which in turn will be used in learning classifiers. We categorize all the features by their *data type* as well as by the **user**(s) from which the features are derived from. By considering different combinations of data types and users, we obtain a comprehensive feature set. By data type, we define: (a) **Content** features: which are derived from textual content of tweets; (b) **Structural** features: which are features derived from the connectivity of target user in relation with his neighbors; (c) **Aggregated** features: which are features derived from summarizing the content or structural features of the target user.

The features can also categorized by the user(s) from which the features are extracted. There are: (a) **Self** features which are extracted from the target user's data only; and (b) **Social** features which are extracted from the neighborhood which includes the target user and other (possibly selected) users connected to him. Assuming that homophily exists, users in the neighborhood are expected to have similar features that enrich the target user representation. Table 4 shows the different combinations of feature categories.

Content Features We consider all original tweets written by the target user excluding the retweets. We call this the user's *tweet document*. Each tweet document is then represented as a bag of words and different content features are constructed from the bag of words. We have considered two types of content features for each word w, namely (a) TF $\times$ IDF, and (b) binary presence of w. TF is the frequency (in log form) of w in the tweet document d ($TF(w,d) = log(1 + f(w,d))$ where $f(w,d)$ is the frequency of word w in d). IDF denotes the inverse document frequency of w ($IDF(w) = log\frac{|D|}{|\{d \in D, w \in d\}|}$. D denotes the set of all user tweet documents. Because the results for TF$\times$IDF consistently outperforms those for binary presence, only TF$\times$IDF results are reported.

Social Content Features. Beside the content features from the user himself, content features of followees are also important for determining user class label. We define a Twitter user's neighborhood to consist of himself and other important users he follows. When a user follows important religious accounts, the additional content features from these accounts will enrich the content of the target user especially for target user who is not actively tweeting.

In our dataset, a user can follow many followees ($min = 0$, $max = 16017$, $average = 290$ for Christians, $min = 0$, $max = 11219$, $average = 309$ for Muslims). We therefore chose for each user top k ($k = 20$) most important followees. We can adopt several different ways of measuring user importance. To obtain social content features, we first combine the target user's tweet document with those tweet

Table 5: Social and Self Content Features (Precision/Recall/F1)

		Christian	Muslim
Self Content only		.792 .880 .834	.844 .775 .808
Social Content (Nghbrs only)	**Degree**	.788 .871 .827	.814 .763 .788
	RInDeg	.792 .869 **.829**	.822 .792 **.807**
Social Content (Nghbrhd)	**Degree**	.815 .897 .854	.879 .810 .843
	RInDeg	.830 .926 **.876**	.901 .837 **.868**

Table 6: All Features using RInDeg as User Importance (Precision/Recall/F1)

		Christian	Muslim
Self	**Content**	.792 .880 .834	.844 .775 .808
	Structure	.690 .986 .812	.960 .426 .590
	Aggregated	.847 .955 .898	.930 .777 .847
Social	**Content**	.830 .926 .876	.901 .837 .868
	Structure	.760 .981 .856	.961 .598 .737
	Aggregated	.769 .954 .852	.924 .728 .814
Self + Social		.851 .976 **.909**	.902 .859 **.880**

documents of his top k followees. The social content features are then extracted from the combined tweet document.

Social Structure Features. Intuitively, a user's choice of followees may reveal her latent attribute. In particular, we assume that religious users tend to follow users important to the religious community and they could be famous pastors, preachers, and religious organizations. We therefore define *social link features* to be derived in two steps: (1) find top k important followees in each religion (2) determine the presence or absence of a follow link to each of the k important followees.

5. EXPERIMENTS

Evaluation on ground truth labels. In the first set of experiments, we evaluate the performance of our prediction method against the 1000+ ground truth labeled users. We use WEKA [5], a machine learning software for computing experiment metrics. Weka includes a wrapper of libSVM, an implementation of SVM by Chang and Lin [4]. Using 10-fold cross validation, we obtain the performance of the classifiers using different combinations of features metrics. Our performance metrics are the standard precision, recall and F1 scores for the Christian and Muslim classes.

Performance of using content features only comparing self content with content generated in the neighborhood is depicted in Tables 5. The results show that social content from neighbors only is comparable to self content and social content from neighborhood outperforms both self content and social content from neighbors only. This observation is consistent with Al Zamal and colleagues' findings [1]. Different from their experiment, our proposed RInDeg outperforms degree in deriving social content for the target users. We shall therefore use RIndeg in the subsequent experiments.

The difference between social networks analysis and traditional text analysis lies in the integration of additional social information. Therefore, we are interested in investigate the effect of integrating additional structural and aggregated features. The results in Tables 6 show that combining both kinds of features indeed improves classification performance. The best F1 scores achieved for Christian and Muslim classes are 0.909 and 0.880 respectively.

As SVM output feature weights, Table 2 shows top 10 induced linguistic terms for two classes. The Christian words

Table 7: Top 10 Selected Terms

Christian	Muslim
psalm, hillsong, corinthians, gospel, proverbs, testimony, church, christ, amen, bible,	allah, insyaallah, masjid, terawih, pakai, awak, pasal, quran, insya, pulak

are indeed related to Christianity. The Muslim words are in Malay because most Singaporean Muslims are ethnically Malay and vice versa.

Evaluation on all users. We next evaluate our classification method on all the 110K+ users who have not been assigned any religion labels. These users do not come with ground truth labels. We therefore manually judged the top 50 scored users under the Christian class, and those top 50 scored users under the Muslim class. The manual judgement was conducted by (1) examining religion related tweets generated by the users, (2) checking their profile pages including biography, recent tweets, and (3) checking his followings and followers. The results show that these top scored users are indeed assigned with the correct religions,i.e., the precision of the manual check is 100%. We also found some non-Malay users (e.g., sirxjcz) who are muslim. While these empirical results are encouraging, we shall conduct a larger scale evaluation on this results in the future work.

6. CONCLUSIONS

As users generate content and follow others on social networks, they leave trace for inferring their latent attributes. Differ from previous works which rely on text features only, we use social links and neighbors to derive useful structure features and social content features. Our proposed classification model with multiple types of features yields F1-score larger than 0.8 for both Christian and Muslim using user provided ground truth data as well as manually judged labeled users.

Acknowledgement

This work is supported by the National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

7. REFERENCES

[1] F. Al Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*, 2012.

[2] S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky. Broadly improving user classification via communication-based name and location clustering on twitter. In *NAACL*, 2013.

[3] A. Bhargava and G. Kondrak. Language identification of names with svms. In *HLT*, 2010.

[4] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[6] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *SIGKDD*, 2011.

[7] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Int'l Workshop on Search & Mining User-generated Contents*.