On Predicting User Affiliations Using Social Features in Online Social Networks

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

On Predicting User Affiliations Using Social Features in Online Social Networks

Minh-Thap Nguyen

SINGAPORE MANAGEMENT UNIVERSITY

2014

On Predicting User Affiliations Using Social Features in Online Social
Networks

by

Minh-Thap Nguyen

Submitted to School of Information Systems in partial fulfillment of the
requirements for the Degree of Master of Science in Information Systems

Dissertation Committee:

Ee-Peng LIM (Supervisor/Chair)
Professor of Information Systems
Singapore Management University

Jing JIANG
Assistant Professor of Information Systems
Singapore Management University

Hady Wirawan LAUW
Assistant Professor of Information Systems
Singapore Management University

Singapore Management University
2014

Abstract

User profiling such as user affiliation prediction in online social network is a challenging task with many important applications in targeted marketing and personalized recommendation. The research task here is to predict some user affiliation attributes that suggest user participation in different social groups.

One of user profiling tasks is religion profiling. Religious belief plays an important role in determining the way people behave, 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 or social variants for representing a user. We introduce the notion of *representative user* to identify users who show religious interest. We further define features using representative users. We first propose a *supervised multiclass classification* method using our proposed features can accurately assign Christian, Muslim, and Buddhist labels to a set of Twitter users with known religion labels.

We further proposed *collective classification* method which make use of additional top classification score unlabeled users. Adding top classification score users provides a means to improve classification accuracy. We present a thorough experiment to show the effective of proposed method.

Acknowledgements

I would like to express my deep appreciation to my supervisor, Professor Ee-Peng Lim for his close guidance. He always gave useful ideas and advices for doing research and running experiments. He clearly showed me the weakness in my writing and gave direction for correction. Without his persistent help, this dissertation will not be possible.

I would like to thank all SMU Professors who did give me wonderful lectures and project guidances. The Professors' enthusiasm presented in high-quality lectures and project meetings greatly inspired me. SMU education will give me great foundation for my career.

I would like to express my gratitude to all SMU colleagues, friends, and staffs for their friendship and helpfulness. They constantly helped me in various aspects of works and life.

I want to dedicate this thesis to my parents for their endless love, support, and encouragement.

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

1.1 Motivation.

In many consumer and social applications, user attributes are required to suggest relevant and interesting products, content, services and social links [22, 37, 3, 8]. Among many user attributes, it is religion which has proven to be very important in determining the way users behave, form preferences, interpret events around them, and develop relationships with others [13, 2, 35, 4]. In the past, users' religion labels are obtained by conducting large scale user surveys run by government agencies and large businesses with or without financial incentives (e.g., lucky draws, direct discounts, etc.) [32]. These large-scale census efforts are generally effective but are also intrusive and time consuming. For online social media users, their religion attributes could well be embedded in the content and their interaction with other users. The question here is therefore how these user religion labels can be recovered from the user-generated data accurately.

1.2 Objectives.

In this thesis, 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 community of 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[1].

User religion prediction task is challenging. Most users are not expected to reveal their religion labels explicitly. Out of the 111,767 users we are able to identify, only 1050 users, or $< 1\%$, 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.

1.3   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

---

[1]We plan to make this dataset publicly available for research purposes.

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 important users among users sharing the same religion. Based on this measure, we derive content and structure features that improve the prediction accuracy.

- We proposed two methods for the multiclass classification problems. The first method make use of labeled data together with a post-training threshold adjustment technique. The second method, namely Collective Classification, iteratively acquires high classification score unlabeled data to train higher accurate classifiers. Both proposed methods yield F1-scores larger than 0.9 for the Christian and Muslim labels, F1-scores larger than 0.7 for Buddhist label. Such an accuracy level makes the classifiers useful in different real world applications.

Chapter 2

Literature Survey

There are quite a number of related works on mining online social media user attributes including political affiliation, gender, ethnicity, and country. Bhargava and Kondrak performed language classification for people names using word and n-gram name features [6]. 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 [24, 25]. 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. Bergsma, Dredze, et al., derived clusters of users and used cluster information to further derive features that can be used for user attribute classification [5]. Their experiments showed that word tokens, n-gram (with $n \leq 4$) and cluster features of user names and locations give more accurate classification of country affiliation, language, ethnicity, gender and race of users. Al Zamal et al. [1] showed that using neighbors' features only to predict age and political affiliation outperform using user features only. This can be attributed to attributes' high assortativity. Rao et al. [27] 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 uses 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.

2.1 Language Classification

Bhargava and Kondrak [6] proposed a model that takes n-gram of names as input and outputs the languages of the names. The model counts the occurrence of n-gram in the names, for n up to a maximum length. The classifier SVM was used because of its ability to handle large number of features and automatically weigh them appropriately. When counting n-gram, the authors included space before and after each word, so that prefixes and suffixed were counted appropriately. In addition to n-gram counts, they also included word length as a feature.

The authors tested three SVM kernels linear, sigmoid, and radial basis function (RBF). They tested maximum n-gram from 1 to 6. In experiments, they doesn't normalize the feature vectors or decreasing the weights of frequent n-gram counts since these didn't bring a considerable improvement.

In the experiments, they used LIBLINEAR package for linear kernel and LIBSVM packet for the RBF and sigmoid kernels. They discarded any periods and parentheses while kept apostrophes and hyphens and convert all letters to lower case. They removed short names of less than two letters. For all data sets, 10% of the data was held out as the test set. They then found optimal parameters for each kernel type by running 10-fold cross validation on the remaining training set. This gave them optimum n-gram lengths of four for single names and five for full names. Using these optimal parameters, they constructed models from the entire training data and tested the models on

the held-out test set. This model outperforms previous language models and linear kernel outperforms other kernels.

2.2 Political Affiliation Classification

One of the early papers about political affiliation classification is Rao et al.'s paper [27]. The paper addressed four problems of Twitter user classification.

1. Political affiliation classification task, there are two labels: Democrats and Republicans.

2. Age classification task, the two labels are below 30 and above 30.

3. Gender classification task, the two labels are males and females.

4. Regional origin classification, the two labels are North and South Indians.

This paper made use of linguistic features only. They proposed two sets of linguistic features:

1. A predefined set of socio-linguistic features (socling). These are important words for classify people from different social groups. These features include emoticons (such as :-), :-(, :—), abbreviations (OMG, LOL), ellipses or puzzled punctuation ("...", !!!, !?!?), exasperation and agreement (Ugh, hmm, ahhh, grrr, yea, yeah).

2. n-gram features including unigrams and bigrams from user linguistic content. Number of n-gram features is several millions. For political affiliation task, number of n-gram features is 4.4 millions while for age classification, number of n-gram features is 4.9 millions, for gender classification number of n-gram features is 1.3 millions.

A model called stacked model which combines the two above mentioned sets of features was included.

The feature values are term frequency of the features in user linguistic content. Each SVM is trained on either socio-linguistic features, n-gram features or both 2 set (stacked model).

In political affiliation task, the accuracy of socio-linguistic model 62.37% is worse than the accuracy of n-gram model 82.84%. Combining both 2 types doesn't improve accuracy 80.19%. Post hoc analyse discovers favorite words used by two classes Democrats and Republicans.

- Democrats like to say about "my youthfull", "my yoga", "my vegetarianism", "my upscale", "my tofurkey", "mysynagogue".

- Republicans like to say about "my zionist", "my yuengling", "my weapons", "my walmart", "my trucker", "my patroit", "my lsu".

It also can be seen that Democrats prefer TV channels such as MSNBC, CNN, NBC, and Logo while Republicans prefer Fox.

Pennacchiotti and Popescu [24] proposed a machine learning approach to classify user political affiliation, ethnicity, Starbucks fans on Twitter.

1. In political affiliation classification task, the two groups to be classified are Democrats and Republican.

2. In ethnicity classification task, the target group is African American.

3. In Starbucks fans classification task, Starbucks fans is the target group to be classified.

The model has an exhaustive and thorough list of features computed from user profile, user tweeting behavior, user linguistic content, and user connected network.

1. The user profile features (PROF-ALL) include the length of user name, number of numeric and alphanumeric characters in user name, use of

profile picture, number of followers, number of followees, followees-to-followers ratio, date of account creation, matching of various regular expression patterns, presence of the location field.

2. The user tweeting behavior features (BEHAV-ALL) include number of tweets posted by the user, number and fraction of tweets that are retweets, number and fraction of tweets that are replies, average number of hashtags and URLs per tweet, fraction of tweets that are truncated, average time and standard deviation between tweets, average number and standard deviation of tweets per day, fraction of tweets posted in each of 24 hours.

3. Linguistic content features (LING-ALL) include prototypical words (LING-WORD), prototypical hashtags (LING-HASH), generic LDA (topics obtained by running LDA on the whole data set of users) (LING-GLDA), domain-specific LDA (topics obtained by running LDA on the domain-specific data set) (LING-DLDA), sentiment words (LING-SENT).

4. The user network features (SOC-ALL) include "friend" (or "followee") accounts (SOC-FRIE), prototypical replied accounts (SOC-REP) and prototypical retweeted accounts (SOC-REP).

One of the contribution of this paper is the proposed way to select relevant features (prototypical words, prototypical hashtags, prototypical followees, prototypical replied accounts, and prototypical retweeted accounts) for each label. Basically, each word is assigned a score equals to number of occurrences in the target label divided by number of occurrences in all labels. By computing this score for every word, they found that the top induced words used by Democrats are "inequality", "homophobia", "woody", and "socialism" while the top induced words used by Republicans "obamacare", "liberty", "taxpayer", and "patriots". By computing this score for hashtags, they found that the top induced words used by Democrats are #itgetsbetter, #VOTE2010,

#ProgCa, and #vote-Dem while the top induced words used by Republicans are #cagop, #ConsNC, #ObamaTVShows, and #RememberNovember. By computing this score for replied users, they found that the top induced users replied by Democrats are txvoodoo, polipaca, liberalcrone, and socratic while the top induced users replied by Replublicans are itsonlywords, glenasbury, RickSmall, and astroterf. By computing this score for retweeted users, the top users retweeted by Democrats are ebertchicago, BarackObama, KeithOl-bermann, and GottaLaff while the top users retweeted by Republicans are Drudge_Report, michellemalkin, fredthompson, and mikepfs. By computing this score for followed users (or followees), the top induced users followed by Democrats are Barack Obama, RachelMaddow, Al Gore, and Keith Ol-bermann while the top induced followed users followed by Republicans are Michelle Malkin, Heritage Foundation, Glenn Beck, and Newt Gringrich.

The two baselines in Pennacchiotti's paper [24] are (B1) classifier that classify user explicitly state their political affiliations and (B2) classifier trained on the profile only. The classifier used in this paper is distributed Gradient Boost Decision Tree (GBDT) run over Hadoop.

B1 has very high precision (0.989 for Democrats, 0.920 for Republicans) but very low recall (0.183 for Democrats, 0.114 for Republicans) and therefore low F-measure (0.308 for Democrats, 0.203 for Republicans). The social features yields ($F1 = 0.896$ for Democrats, $F1 = 0.796$ for Republicans) better results than the linguistic features ($F1 = 0.825$ for Democrats, $F1 = 0.668$ for Republicans), and they both work better than the baselines B1 ($F1 = 0.308$ for Democrats, $F1 = 0.203$ for Republicans) and B2 ($F1 = 0.808$ for Democrats, $F1 = 0.533$ for Republicans). Together the model of all features yields best performance ($F1 = 0.910$ for Democrats, $F1 = 0.833$ for Republicans).

Another approach was proposed by Al Zamal et al. [1]. The authors considered three tasks:

1. Political affiliation classification of Democrats and Republicans.

2. Age classification of 18+ (age from 18 to 23) and 25+ (age from 25 to 30).

3. Gender classification of males and females.

First, linguistic, tweeting behavior, and network features are derived for all users.

Here are the list of features:

- k-top words,

- k-top stems,

- k-top digrams and trigrams,

- k-top co-stems,

- k-top hashtags,

- frequency statistics: tweets, mentions, hashtags, links, and retweets per day,

- retweeting tendency: retweets-to-tweets ratio,

- neighborhood size: followers-to-followees ratio.

This paper chooses top words in the same way of choosing prototypical words in Pennacchiotti and Popescu's paper.

The authors of this paper do not make use of social network features but make use of the features of target user's neighborhood by aggregating the features of target user's neighbors to the features of target user. They define neighborhood policies to choose for each user a set of appropriate neighbors in order to test which subset of neighbors will yield the best performance.

1. All neighbors,

2. n-most popular neighbors based on number of followers,

3. n-least popular neighbors,

4. n-closest neighbors who receive the most number of mentions.

Their finding is quite interesting: for political affiliation classification task which has the highest homophily, integrating features from all neighbors improve the classification results the most with the accuracy improved from 0.890 to 0.932.

Rao et al. [27], Pennacchiotti and Popescu [24], and Al Zamal et al. [1] address political affiliation classification. However, Rao et al. used linguistic features only while Pennachiotti and Popescu used profile features, tweeting behavior features, prototypical linguistic content features, prototypical social network features; Al Zamal et al. used k-top words, k-top stems, k-top digrams and trigrams, k-top co-stems, k-top hashtags, frequency statistics, retweeting tendency, neighborhood size. The difference between Pennacchiotti's model and Al Zamal's model is that Pennacchiotti used prototypical social network features while Al Zamal used social network information to update user's features using neighborhood's features.

Political affiliation task is the task that both three above-mentioned papers consider two labels Democrats and Republicans with a balanced data set. The accuracy of the task is reported by both three papers as following.

- Rao et al. gained an accuracy of 82.84%. by using n-gram features only;

- Pennacchiotti and Popescu obtained an accuracy of 88.3% on user features only and 88.9% on graph-updated results;

- Al Zamal et al. acquired an accuracy of 89.0% on user features only and 93.2% on neighborhood features

2.3   Age Classification

Both Rao et al. [27] and Al Zamal et al. [1] classified age but they differ in how they label: Rao et al. considered below 30 and above 30 while Al Zamal considered 18-23 vs. 25-30. Another difference is Rao et al. used linguistic features only while Al Zamal used selected linguistic features and frequency statistics, retweeting tendency, neighborhood size.

In Rao's paper, socio-linguistic model performs worse than n-gram model (Accuracy 69.44 compares to 73.09). The predefined set of features doesn't general well compares to full set of features. For above 30 year old users, the distinctive words used by them are "my work", "my epidural", "my daughters", "my grandkids". While for below 30 year old users, the distinctive words used by them are "my zunehd", "my yuppie", "my sorors", "my rents", "my classes".

In Al Zamal's paper, aggregating features from neighbors does improve classification results compare to using user self features only (Accuracy 0.751) and aggregating features from least important followees yields the best results (Accuracy 0.782). We can infer that the the neighbors who have least number of followers are the ones who share the same age with the target user. This characteristic is call homophily.

2.4   Gender Classification

Rao et al. [27] also works on gender classification. In this task, using predefined set of socio-linguistic words performs better than using complete set of n-grams: accuracy 71.76% compares to 68.70%. This infers that the different genders do differ in the way they use socio-linguistic words. The most different features are emoticons (females use 3.5 times more than males), ellipses (...) or repeated exclamations (females use 1.5 times more than males). Post hoc analyse discovers favorite words used by two classes Males and Females.

- Male like to say about "my zipper", "my wife", "my gf", "my nigga", "my want".

- Female like to say about "my zzz", "my yoghurt", "my yoga", "my husband", "my bf".

Al Zamal et al. [1] used more selected set of linguistic features aggregated with features from neighbors. They have shown that integrating social features does not improve performance in this task. That can be attributed to the fact that homophily property does not apply to gender.

2.5 Summary

Both three papers by Rao et al. [27], by Pennacchiotti and Popescu [24], and by Al Zamal et al. [1] each one proposed a single model to solve several Twitter user classification tasks. Rao et al,'s model included only linguistic content features which comprised of predefined socio-linguistic features and n-gram features. Pennacchiotti and Popescu proposed a thorough set of features including both linguistic features and social features but not social linguistic features. Al Zamal et al. proposed inclusion of features from neighbors to improve overall classification.

The main findings of Pennacchiotti are

- Method to compute score and rank features for each class.

- In all three tasks political affiliation, ethnicity, Starbucks fan, including social features help to improve classification results.

- Graph-based score update that update user classification score based on his neighbors' scores. This works for political affiliation classification task.

The main findings of Al Zamal are

- Method to include features from neighborhood. This works for homophily attributes like political affiliation and age classification.

- Different classification tasks have different optimized set of neighbors: for political affiliation it is all neighbors, for age classification it is least important neighbors, for gender classification it is closest neighbors.

What is common in these paper is the finding political affiliation has homophily attributes as graph-based score update [24] and neighborhood feature update [1] help to improve results. Other tasks like ethnicity classification, Starbucks fans classification, gender classification do not show this property.

Age possess homophily attribute because friends are usually of similar ages. Age's homophily attribute is stronger than gender's but weaker than political affiliation's is an observation by Al Zamal et al. [1]. Their proposed model yield higher improvement for political affiliation than age when integrating features from neighbors.

Gender differ from political affiliation and age in a way that it doesn't have homophily attribute meaning if two users connect to each other doesn't mean they will likely share the same gender. This will have the effect that aggregate feature from neighbors will not help to boost classification performance. Nevertheless, different genders have different follow habits. Therefore, if we find these differences and aggregate them as features, we can improve classification results.

The shortcoming of Pennacchiotti's method is that it only considers each user's out-link as a feature, we can extend to consider user's structure features. For example, we can use number of out-links to users of each classes or the class of that the user has the maximum number of out-links to users.

The shortcoming of Al Zamal's method is their way of ranking neighbors merely based on the degrees. For example, they choose neighbors who have the most number of followers or the least number of followers. Degree is only a measure of influence not specific to any label. This should be changed to a

degree that takes into account the different classes, for example, the ratio of the number of followers of two classes.

Chapter 3

Religion Prediction using Social Network Data

In this section, we define the religion prediction task, together with our proposed method, multiclass classification using content and social structural features, and the experiments to verify our proposed method. Our method adopts a taxonomy of features which can be extracted from the Twitter data. The performance of our method is evaluated against several baselines using a specially constructed labeled dataset.

3.1 Religion Prediction Task

This research is conducted on the users of Twitter, a popular microblogging site. Each Twitter user can optionally provide a written biography covering his/her interests and other profile information.

As Twitter becomes highly popular among the social media users. It has also attracted businesses promoting and sensing feedbacks to their products and services. Twitter is also used by government agencies for engaging citizens, by socio-political groups for promoting events and causes, as well as by religion organizations for propagating beliefs.

As a Twitter user, one can

1. share thoughts, current activities, and URLs in a 140 character limit messages also known as "tweets";

2. repost (or retweets) other users' tweets;

3. communicate with other users by sending and receiving private messages;

4. follow other users to receive their updates.

The **religion prediction task** for a set of Twitter users is one that assigns religion label(s) to every user using all user-generated data. In this thesis, we perform this task on a set of Singapore Twitter users. Hence, the set of religion labels considered are the main religions found in Singapore, namely, Christianity, Islam, Buddhism, and others.

Religion is largely a permanent user attribute. Users normally do not change their religion beliefs. In some case, abandoning a religion could be considered a crime and people doing so would be punished and deserted by their communities. Religion is closely related to culture and ethnicity. Therefore, believers of the same religion tend to share common languages and habits. Religion also bring people together. People are therefore more likely to develop connections with others with the same religion. All these characteristics of religion make the study of religion prediction using Twitter user network, content and other data highly interesting.

We propose two methods to this religion prediction task: the **Multiclass Classification** method which uses content and social structural features and the **Collective Classification** method [21, 30]. Both methods make use of a rich set of features derived using a feature taxonomy developed by this thesis.

Both our proposed methods follow a series of steps: (1) Feature Engineering; (2) Classifier Training using Labeled Data; (3) Classification on Unlabeled Data; and (4) Performance Evaluation. We will begin by constructing a Twitter dataset with ground truth religion labeled users including Christians, Muslims, and Buddhists. This ground truth dataset construction is presented in

Section 3.2. After that, we will derive feature vector of every user. A classifier will be learned from the feature vectors of labeled users and then be applied to unlabeled users. Then, top confidence scored users are added as pseudo-labeled users because they are deemed to improve classification performance.

3.2 Ground Truth Label Assignment

Construction of Twitter Dataset. We crawled the Twitter data generated by about 110,000 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 find other Singapore user accounts connected to the seeds. The process is repeated several times before reaching the above user count. The dataset consists of users' profiles, tweets, and follow links among the users. Table 3.1 shows the basic statistics of this dataset.

Table 3.1: Singapore Twitter Dataset

# total users	111,767
# total follow links	1,770,272
# labeled as Christian	581
# labeled as Muslim	448
# labeled as Buddhist	21

Table 3.2: Selected keywords for assigning religion labels

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

Table 3.3: Composition of Religions in Singapore (2010)

Buddhism	33.3%
Christianity	18.3%
Islam	14.7%
Taoism	10.9%
Hinduism	5.1%
Other Religions	0.7%
Non-religious	17.0%

User religion labeling. There are five main religions in Singapore, namely, Buddhism, Christianity, Islam, Taoism, and Hinduism. We searched religion specific keywords in the biography of users as shown in Table 3.2 these keywords include 'jesus', 'christ' for Christians, 'allah' for Muslims, and 'buddha' for Buddhists. For each user with biography containing above keywords, we manually judged the religion affiliation of the user based on the biography content. Only those who clearly state their religion affiliation in the biography were then assigned the corresponding ground truth religious labels. As shown in Table 3.1, we managed to label 581 Christians, 448 Muslims, and 21 Buddhists. This religion composition is quite different from that reported in the 2010 Singapore population census (the most recent census)[32] shown in Table 3.3. In that table, Buddhists form 33.3% of the population followed by 18.3% Christians and 14.7% Muslims. This suggests that Christian and Muslim users have higher propensity to share their religion beliefs online.

Representative Users. Among the users in the 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 instead interested in measures based on the user's importance in each religion group. We therefore define two new measures, called *label-indegree* and *representative indegree ratio*, as follows.

- The **label-indegree** of user u for label r, $indeg(u, r)$, is defined as the number of in-links to user u from users with religion label r.

- The **representative indegree ratio** of u for label r, $RInDeg(u, r)$, is defined as: $RInDeg(u, r) = \frac{indeg(u,r)+\alpha}{indeg(u,r'_{max}(u))+\alpha}$ if $r = arg\max_{r'} indeg(u, r')$ and $RInDeg(u, r) = 0$ otherwise, where $r'_{max}(u)$ return the religion label of the next largest religion group among followers of u ($r'_{max}(u) = $

$arg \max_{r' \neq r} \text{indeg}(u, r')$). 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 indegree ratio $\text{RInDeg}(u, r)$ (≥ 1) measures the dominance of a specific religion compared with other religions among u's followers. The larger $\text{RInDeg}(u, r)$, the more dominance is the religion r among followers of u. This suggests that u is very important among the followers with the religion r. When u is followed by only users of a single religion, $\text{RInDeg}(u, r) = \frac{\text{indeg}(u,r)}{\alpha} + 1$. For $\alpha = 1$, $\text{RInDeg}(u, r) = \text{indeg}(u, r) + 1$. When u is not followed by any users with religion label or when u is followed by equal sized religious user groups, $\text{RInDeg}(u, r) = 1$. Figure 3.1 illustrates differences of the above two measures using an example.

User u has 4 Christian followers c_1, c_2, c_3, c_4, 3 Muslim followers m_1, m_2, m_3, and 1 Buddhist follower b_1
Therefore,
```
indeg(u,Christian) = 4
indeg(u,Muslim) = 3
indeg(u,Buddhist) = 1
r_max(u) = Christian
r'_max(u) = Muslim
RInDeg(u,Christian) = 1.25
RInDeg(u,Muslim) = 0
RInDeg(u,Buddhist) = 0
```
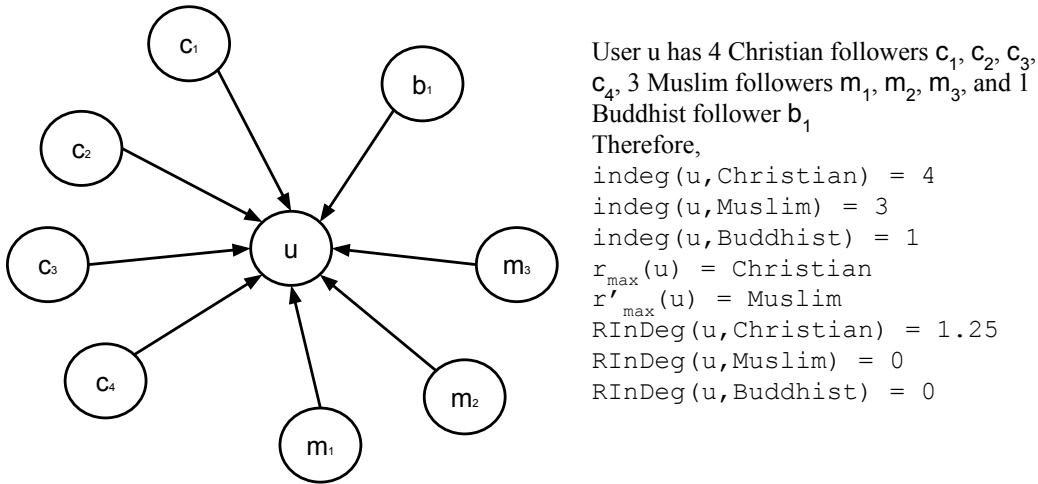
Figure 3.1: Label-Indeg and RInDeg

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

We apply the two proposed representativeness measures on two sets of users: Singapore users and non-Singapore users. The reason we have to consider non-Singapore followees is that they may be actively followed by Singa-

pore users. These non-Singapore followees may provide useful features for our prediction task.

Table 3.4: Top Singapore Representative Users By Label-Indegree

Rank	indeg(u, r)		
	r=Christian	r=Muslim	r=Buddhist
1	konghee (139)	TaufikBatisah (126)	mrbrown (6)
2	STcom (127)	SoSingaporean (97)	STcom (6)
3	JosephPrince (90)	banchothematrep (93)	SoSingaporean (5)
4	mrbrown (87)	sezairi (69)	miyagi (4)
5	SoSingaporean (82)	WaktuSolatSG (69)	iamnatho (4)
6	chcsg (63)	STcom (66)	Xiaxue (4)
7	ChannelNewsAsia (52)	FauzieLaily (62)	LeticiaBongnino (4)
8	LeticiaBongnino (49)	HadyMirzaHM (59)	imMichelleChong (3)
9	nccsg (44)	Norfasarie (58)	wpsg (3)
10	Xiaxue (44)	ChannelNewsAsia (43)	YahooSG (3)

Table 3.5: Top Singapore Representative Users By RInDeg

Rank	RInDeg(u, r)		
	r=Christian	r=Muslim	r=Buddhist
1	konghee (140)	Norfasarie (59)	jolantru (4)
2	JosephPrince (91)	MediaCorp_Suria (32)	NESociety (3)
3	chcsg (64)	iNasuha (29)	visakanv (3)
4	nccsg (45)	fadhilfadaroz (26)	trueboat (3)
5	joepurcell (36)	didicazli (25)	sg_kopitiam (3)
6	Celestfoo (35)	MuslimSG (24)	WriteClique (2)
7	thezoneministry (33)	HyrulAnuar (24)	wayangparty (2)
8	JianMingTan (27)	DearAbdullah (23)	publichousesg (2)
9	garrettleejw (27)	Fiza_O_ (23)	hai_ren (2)
10	Chris_Honegger (25)	TaufikBatisah (21)	CJrystal (2)

Table 3.6: Top non-Singapore Representative Users By Label-Indegree

Rank	indeg(u, r)		
	r=Christian	r=Muslim	r=Buddhist
1	RickWarren (129)	IslamicThinking (87)	DalaiLama (10)
2	JoyceMeyer (100)	IslamSpeaks (82)	tinybuddha (7)
3	MaxLucado (90)	LisaSurihani (53)	BarackObama (6)
4	DarleneZschech (87)	Retwittings (48)	BillGates (6)
5	JohnBevere (84)	justinbieber (46)	Zen_Moments (6)
6	CSLewisDaily (78)	NaomiNeo_ (45)	taylorswift13 (5)
7	JohnCMaxwell (78)	MaherZain (45)	TheEconomist (5)
8	philpringle (76)	TheEllenShow (43)	HarvardBiz (5)
9	ARBernard (76)	TheNobleQuran (42)	dailyzen (5)
10	JoelOsteen (71)	zaynmalik (38)	Buddhism_Now (5)

Table 3.7: Top non-Singapore Representative Users By RInDeg

Rank	RInDeg(u, r)		
	r=Christian	r=Muslim	r=Buddhist
1	JoyceMeyer (101)	IslamicThinking (88)	Zen_Moments (7)
2	MaxLucado (91)	IslamSpeaks (83)	Buddhism_Now (6)
3	DarleneZschech (88)	LisaSurihani (54)	elephantjournal (6)
4	JohnBevere (85)	MaherZain (46)	DhammaLinks (5)
5	CSLewisDaily (77)	TheNobleQuran (43)	theworsthorse (5)
6	philpringle (77)	syarifsleeq (36)	Bodhipaksa (5)
7	ARBernard (72)	awalashaari (36)	BuddhistGeeks (5)
8	JoelOsteen (68)	Aaron535Aziz (36)	DharmaDots (5)
9	hillsongunited (65)	WardinaSafiyyah (33)	waylonlewis (5)
10	BrianCHouston (60)	ImranAjmain (32)	tricyclemag (5)

Tables 3.4 and 3.5 list the top ten Singapore users for each measure (with measure values given in parentheses). We can see that there are overlapping top representative users across religions by label-indegree and they include *STcom*, *mrbrown*, *SoSingaporean*, and *ChannelNewsAsia* which are owned by popular bloggers, and news agencies in Singapore. RInDeg does well in recognizing the top Singapore representative users unique to each religion, and these user accounts include: *JosephPrince*, *konghee*, *chcsg*, and *nccsg* which belong to popular church leaders and churches, *FauzieLaily*, *Norfasarie*, and *MediaCorp_Suria* which are popular Muslim celebrities and news agencies. All these users are based in Singapore. This observation indicates that RInDeg is better than label-indegree in identifying important persons in the religion communities.

We however realize that the top Singapore users followed by Buddhists are not related to Buddhism. These top RInDeg Singapore users are *mrBrown*, *STcom*, *SoSingaporean*, and *jolanchu*. We therefore broaden our representative users to include non-Singapore followees. As shown in Table 3.6 and 3.7, we found that Singapore Buddhists do follow non-Singapore Buddhists such as *Zen_Movement*, *Buddhism_Now*, and *elephantjournal*. Non-Singapore Christian accounts such as *JoyceMayer*, *MaxLucard*, and *DarleneZschech* are also followed by many Singapore Christian users. Non-Singapore Muslim accounts

such as *IslamicThinking*, *IslamSpeaks*, and *LisaSurihani* (Muslim tweeters) are followed by many Singapore Muslim users.

3.3 Feature Engineering

Twitter data is rich as it contains user attributes, text, and network data. As shown in Figure 3.2 and Figure 3.8, a taxonomy of features is proposed in this work. At the first level of the taxonomy, we categorize features into (1) *Self*; and (2) *Social* features.

Self features are further categorized into (a) *Self-Name*, (b) *Self-Content*, (c) *Self-Structural*, and (d) *Self-Aggregated* features. The Self-Name features consist of character-level n-grams of user full name. The Self-Content features are derived from textual content of tweets. The Self-Structural features are RInDeg, label-indeg, label-outdeg, label-Nindeg, and label-Noutdeg for each religion r of the target user. The Self-Aggregated features are derived from summarizing the Self-Content and Self-Structural features of the target user.

Similarly, the sub-categories of Social features are (a) *Social-Content*, (b) *Social-Structural*, and (c) *Social-Aggregated* features. The Social-Content features include the Social-Neighbor Only features and the Social-Neighborhood features. The Social-Neighbor Only features are derived from textual content of the target user's top k neighbors. The Social-Neighborhood features are derived from the textual content of both the target user and his top k neighbors. The Social-Structural features are the top k $\langle importance \rangle$ neighbors where $\langle importance \rangle$ is measured by RInDeg or label-indeg for each religion r. The Social-Aggregated features are derived from summarizing the social-content and social-structural features.

We also consider the features that are *label-independent* and *label-dependent*. Label-independent features are ones that can be computed without labeled data. Examples of such features include Self-Name and Self-Content features.

The other features are label-dependent.
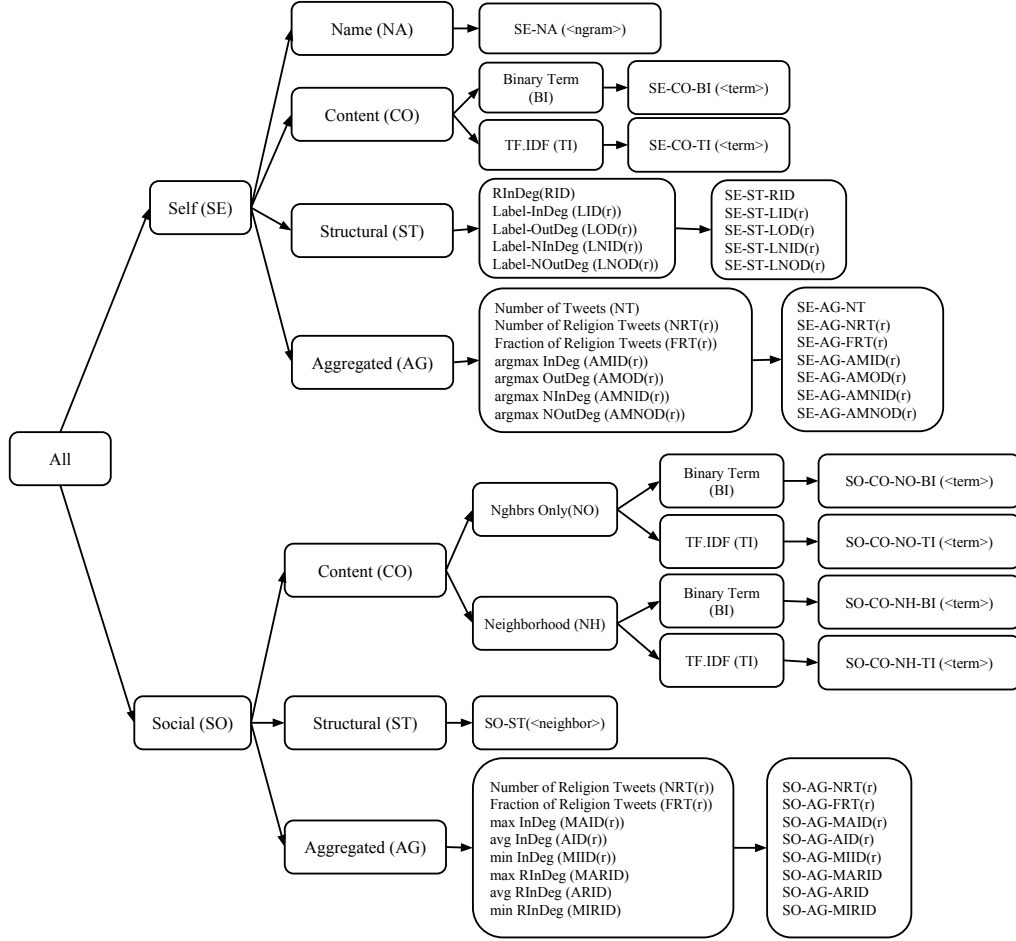
Figure 3.2: Feature Categorization Tree

3.3.1 Self-Name Features

Self-Name features are user full name's character n-grams ($2 \leq n \leq 5$). Assuming that Christian, Muslim, and Buddhist names has different distributions of character n-grams, these n-grams are expected to improve the prediction accuracy.

3.3.2 Self-Content Features

We consider all original tweets written by the target user and exclude the retweets. These original tweets of the user are combined into one tweet document. Each tweet document d is then represented as a bag of words and differ-

Table 3.8: Categorization of Features

	Name (NA)	Content (CO)	Structural (ST)	Aggregated (AG)
Self (SE)	n-grams SE-NA($\langle n\text{-gram}\rangle$)	- binary terms SE-CO-BI($\langle term\rangle$) - TF · IDF SE-CO-TF($\langle term\rangle$)	- RInDeg (SE-ST-RID) - label-indeg for each religion r (SE-ST-LID(r)) - label-outdeg for each religion r (SE-ST-LOD(r)) - label-Nindeg for each religion r (SE-ST-LNID(r)) - label-Noutdeg for each religion r (SE-ST-LNOD(r))	- number of tweets (SE-AG-NT) - number/fraction of tweets containing selected keywords for each religion r (SE-AG-NRT(r)/SE-AG-FRT(r)) - $argmax_r$ indeg (SE-AG-AMID(r)) - $argmax_r$ outdeg (SE-AG-AMOD(r)) - $argmax_r$ Nindeg (SE-AG-AMNID(r)) - $argmax_r$ Noutdeg (SE-AG-AMNOD(r))
Social (SO)		- binary terms from neighbors only SO-NO-BI($\langle term\rangle$) - TF · IDF from neighbors only SO-NO-TI($\langle term\rangle$) - binary terms from neighborhood SO-NH-BI($\langle term\rangle$) - TF · IDF from neighborhood SO-NH-TI($\langle term\rangle$)	- top k $\langle importance\rangle$ neighbors $\langle importance\rangle =$ {RInDeg, label-indeg for each religion r } SO-ST($\langle neighbor\rangle$)	- number/fraction of tweets from neighborhood containing selected keywords for each religion r (SO-AG-NRT(r)/SO-AG-FRT(r)) - avg/max/min of $\langle importance\rangle$ of top k $\langle importance\rangle$ neighbors $\langle importance\rangle =$ { RInDeg, indeg for each religion r } (SO-AG-ARID, SO-AG-AID(r), SO-AG-MARID, SO-AG-MAID(r), SO-AG-MIRID, SO-AG-MIID(r))

ent content features are constructed from the bag of words. We have considered two types of content features for each word w, namely (a) $TF(w, d) \times IDF(w)$, and (b) $I(TF(w, d))$. $TF(w, d)$ 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. $IDF(w)$ denotes the inverse document frequency of w: $IDF(w) = \log \frac{|D|}{|\{d \in D, w \in d\}|}$, where D denotes the set of tweet documents of users in training data. Function $I(x)$ returns 1 if $x > 0$ and returns 0 otherwise. Therefore, $I(TF(w, d))$ returns 1 if w is in d and returns 0 otherwise. Self-Content features are label-independent.

3.3.3 Self-Structural Features

Self-Structural features include user's RInDeg, label-indegree, label-outdegree, label-Nindegree , label-Noutdegree (label-outdegree divide by number of users assigned the label) for each religion r. Self-Structural features are label-dependent.

RInDeg and label-indegree for each religion r are defined in the previous section. Here, we also include the other variants for completeness. A user's label-outdegree for each religion r is the number of her followees assigned the label r. A user's label-Nindegree for each religion r is his label-indegree for the same religion divide by the number of users assigned the religion label. Similarly, a user's label-Noutdegree for each religion r is his label-outdegree for the same religion divide by the number of users assigned the religion label.

3.3.4 Self-Aggregated Features

By summarizing Self-Content and Structural features of a user u, we derive the Self-Aggregated features. These features include number of tweets generated by each user, number and fraction of tweets that containing selected keywords for each religion r, $arg \max_r indeg(u, r)$, $arg \max_r outdeg(u, r)$, $arg \max_r Nindeg(u, r)$, and $arg \max_r Noutdeg(u, r)$. For example, in Figure 3.1,

we have $indeg(u, Christian) = 4$, $indeg(u, Muslim) = 4$, and $indeg(u, Buddhist) = 1$, therefore $arg\max_r indeg(u, r)$ is Christian. Self-Aggregated features are label-dependent.

3.3.5 Social-Content Features

Beside the content features from the user himself, the content features of followees provide useful features for determining the class label. We define a user's neighborhood to consist of himself and other important users he follows. When a user follows important religion accounts, the additional content features from these accounts will enrich the content of the target user especially for a target user who does not tweet actively. Social-Content features consist of content features from the neighbors only or content features from the neighborhood including both the user and her neighbors. Social-Content features are label-dependent.

In our dataset, a user can follow many followees (seen Table 3.9). We therefore chose for each user top k ($k = 100$) most important representative users. We can adopt several different ways of measuring user importance.

Table 3.9: Statistics of Users' Followee Counts in Our Data

Religion	Min	Average	Max
Christian	0	290	16,017
Muslim	0	309	11,219
Buddhist	45	387	1,857

3.3.6 Social-Structural Features

Intuitively, a user's choice of followees may reveal her latent attributes. In particular, we assume that religious users tend to follow representative users of that religion who are important to the religious community and they could be famous pastors, preachers, and religious organizations. We therefore define the Social-Structural features to be derived in two steps: (1) find the top k

representative users in each religion; and (2) determine the presence or absence of a follow link to each of these k representative users. Social-Structural features are label-dependent.

3.3.7 Social-Aggregated Features

By summarizing Social-Content and Social-Structural features, we derive Social-Aggregated features. This feature type includes: number and fraction of tweets from neighborhood that containing selected keywords for each religion r, min/max/avg of RInDeg and Label-Indegree of top k representative users. Social aggregated features are label-dependent.

The process to derive all the features is presented in Figure 3.3 in which Steps 1 to 4 are for deriving Self-Name, Self-Content, Self-Structural, and Self-Aggregated features respectively, Steps 5 to 7 are for deriving Social-Content, Social-Structural, and Social-Aggregated features.
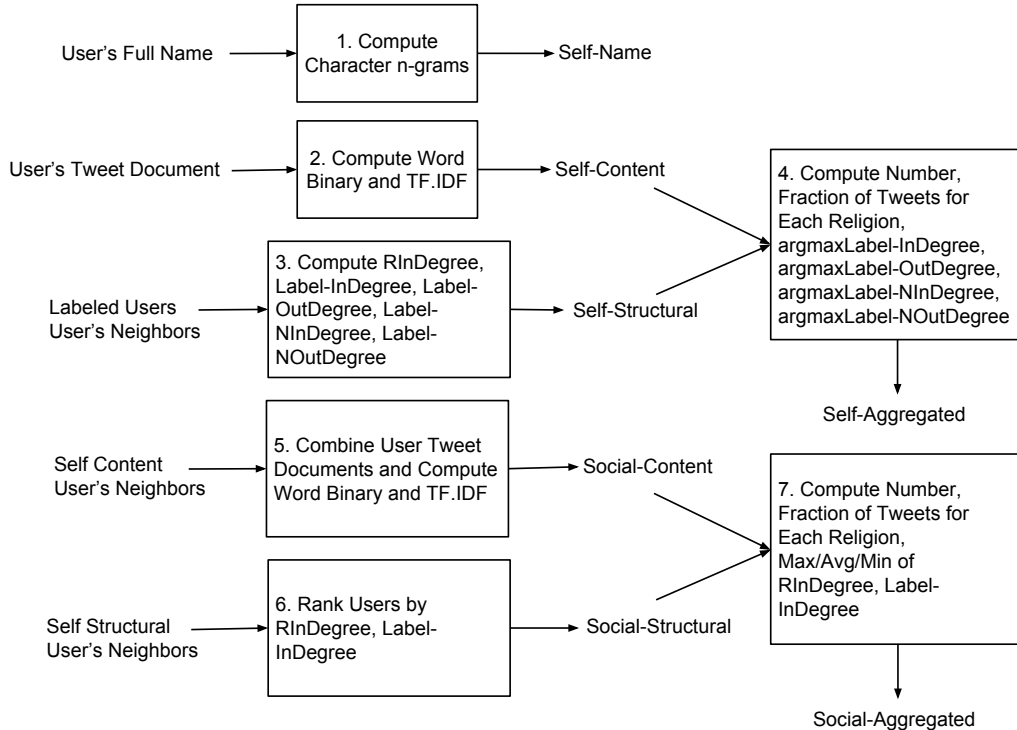
Figure 3.3: Twitter User Classification Design

3.4 Multiclass Classification using Content and Social Structural Features

In our proposed model, we make use of two types of classifiers: Naive Bayes and linear SVM.

3.4.1 Naive Bayes

Naive Bayes [20, 19] is a classifier based on Bayes' theorem. Let consider the supervised learning problem in which we want to estimate an unknown target probability $P(Y|X)$ where Y is the target class and X is the n-dimensional feature vector. The idea of Naive Bayes is to assume that the features $X_1 \ldots X_n$ are conditionally independent of one another, given the target class Y. This leads to the computation of probability $P(X|Y)$ as

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i|Y)$$

The probability that Y will take on a possible value y_j is

$$P(Y = y_j|X_1 \ldots X_n) = \frac{P(Y = y_j)P(X_1 \ldots X_n|Y = y_j)}{\sum_k P(Y = y_k)P(X_1 \ldots X_n|Y = y_k)}$$

Because $X_1 \ldots X_n$ are conditionally independent given Y, we can rewrite this equation to

$$P(Y = y_j|X_1 \ldots X_n) = \frac{P(Y = y_j)\prod_i P(X_i|Y = y_j)}{\sum_k P(Y = y_k)\prod_i P(X_i|Y = y_k)}$$

The denominator is the same for all value y_j. Therefore, we have the Naive Bayes classification rule:

$$Y \leftarrow arg \max_{y_j} P(Y = y_j) \prod_i P(X_i|Y = y_j)$$

Naive Bayes is commonly used in classification task [18] due to its simplicity

and training efficiency. However, its performance is often degraded when the conditional independence does not hold.

3.4.2 Linear SVM

Linear SVM is proven to be excellent in traditional text mining task [15]. A linear SVM [36, 17] is trained to find a hyperplane that maximize the margin between the two classes while accepting and penalizing some misclassified cases. Given a training data $D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^n, y_i \in \{0, 1\}\}$. Linear SVM can be defined as an optimization problem

$$arg \min_{(\mathbf{w},b)} \frac{1}{2} ||\mathbf{w}||^2 + C \sum_i \xi_i$$

subject to

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0, \forall i$$

The hyperplanes are specified by a subset of positive and negative training examples known as positive and negative support vectors (SV) respectively (see Figure 3.4).

Once $\mathbf{w}$ and b are learned, SVM compute each unlabeled instance a score by applying decision function on its feature vector $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b$$

The sign of $f(\mathbf{x})$ is used to predict the instance's label: the instance is labeled positive if $f(\mathbf{x}) \geq 0$ and negative otherwise. In other words, SVM takes 0 as the default threshold, denoted as θ, in its decision function.

Choosing Threshold θ. The main idea of this technique is to change SVM's default decision boundary corresponding to the default threshold $\theta = 0$ to the best decision boundary by choosing the best threshold θ.
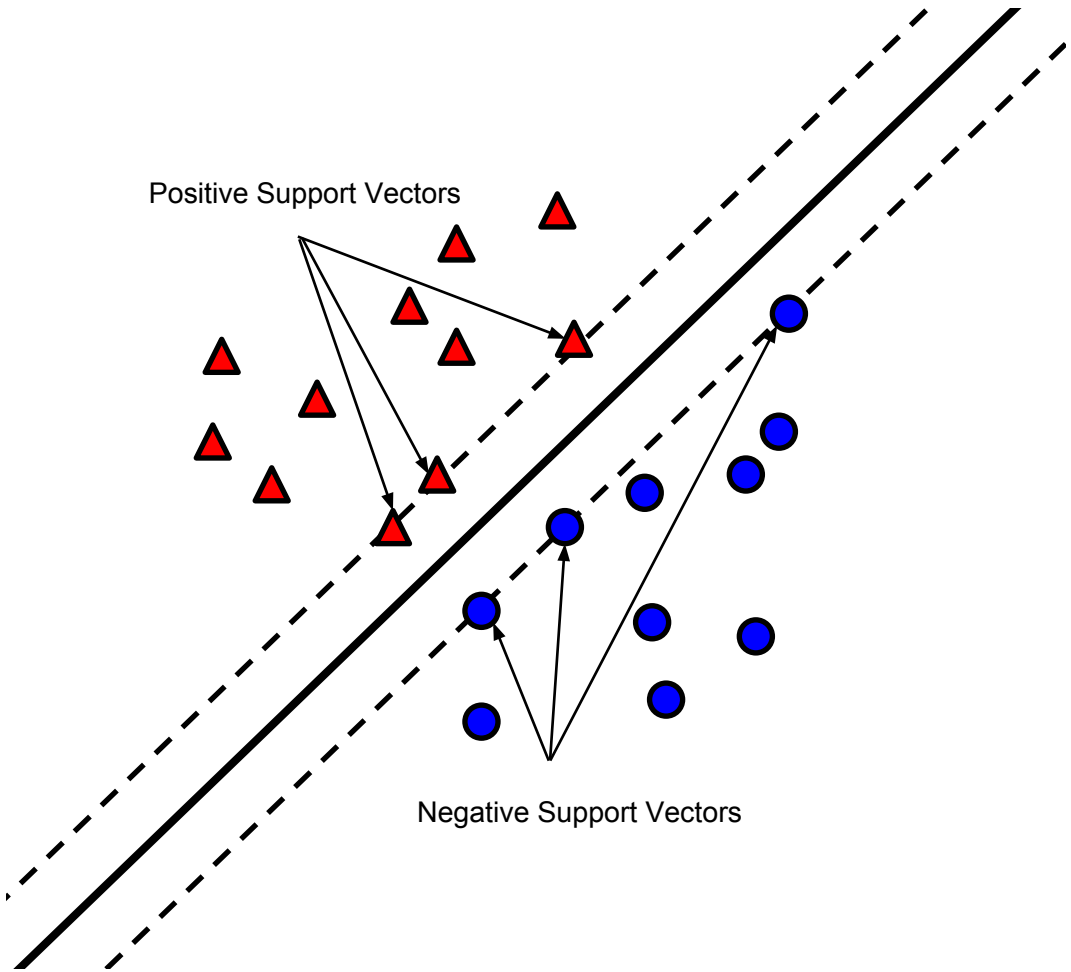
Figure 3.4: Linear Kernel Support Vector Machine

To choose the best threshold θ, all validation instances are ranked in descending order according to their SVM output confidence scores. The top n ranked instances are labeled as positive such that F_1 of the category is maximized. The score of nth instance, denoted as θ, is the optimal threshold. The test instances having confidence scores greater than or equal θ are assigned positive label. When the positive class is small, a negative threshold could be used to accept some negative scored instances [34].

In our religion prediction task, we have very few Buddhists in contrast with large number of Christians and Muslims, making our task imbalanced. The positive class in this setting is the smaller class. In order to mitigate the effect of class imbalance, threshold method that adjusts decision thresholds of a classifier to balance the precision and recall and improve F_1 is employed

[31]. Sun et al. [33] further showed that threshold method alone rather than other methods like resampling or instance weighting help improving SVM's performance.

Threshold selection technique is a post-training strategy as it is applied after training phase of the classifier. It adjusts decision thresholds that accept instances as member of a class [29]. Provost showed that classifiers learned from imbalanced data should apply threshold method [26].

3.4.3 Multi-class Classification

We adopt *one-against-all* learning strategy [16]: train three binary classifiers on labeled data, each for one religion versus the two others, i.e., Christian vs. the rest, Muslim vs. the rest, Buddhist vs. the rest. Then, we apply three classifiers on unlabeled data to compute three scores for each user. In case of $\theta = 0$, a user will be classified as member of religion r ($r \in \{Christian, Muslim, Buddhist\}$) if his score output by the classifier corresponding to r is positive and the largest among three religious scores (if the largest score among three scores is negative then the user will be classified as *Unknown*). In case of $\theta \neq 0$, a user's three scores will be subtracted by the corresponding θ and then the user will be classified by comparing the three modified scores with each other and with 0 in the same way with the case $\theta = 0$.

By doing this, there are still users who do not have any religion label when both their modified scores ($score - \theta$) less than 0. These users are assigned *Unknown* religion label. Users who do not follow any religion, do not tweet about religion or follow any religion-related users should be classified into this label.

As the number of labels increase, the difficulty of the multi-class classification also increase, and therefore need the larger size of the training set. In the multi-class classification, some classes will appear more difficult than others to

classify. The reasons are (1) very few positive training examples for the class; and (2) lack of good predictive features for that class.

3.5 Experiments

3.5.1 Classification Metrics

Van Rijsbergen [28] specified evaluation metrics for evaluating classification performance *precision, recall*, and F_β score (F_β measure).

Metrics for a single fold

Precision for a label c, denoted as Pr_c, measures the percentage of correct assignments among all the instances assigned to label c. Recall for a label c, denoted as Re_c, measures the percentage of correct assignments among all instances which actually have label c. Let TP_c be the set of instances correctly assigned label c, FP_c be the set of instances incorrectly assigned label c, TN_c be the set of instances correctly rejected label c, FN_c be the set of instances incorrectly rejected label c. Precision and recall are defined as follows[1].

$$Pr_c = \frac{|TP_c|}{|TP_c| + |FP_c|}$$

$$Re_c = \frac{|TP_c|}{|TP_c| + |FN_c|}$$

Neither precision nor recall can effectively measure classification performance in isolation [29]. Therefore, the performance has often been measured by the combination of the two measures. F_β measure was proposed to combine precision and recall.

$$F_{\beta c} = \frac{(\beta^2 + 1)Pr_c Re_c}{\beta^2 Pr_c + Re_c} = \frac{(\beta^2 + 1)|TP_c|}{(\beta^2 + 1)|TP_c| + \beta^2|FP_c| + |FN_c|}$$

---

[1]for any set S, $|S|$ denotes the number of elements in S

Normally, F_1 ($\beta = 1$) (the harmonic average of precision and recall) is used.

$$F_{1c} = \frac{2 Pr_c Re_c}{Pr_c + Re_c} = \frac{2|TP_c|}{2|TP_c| + |FP_c| + |FN_c|}$$

Metrics for n-fold cross-validation

Forman and Scholz [11] suggested how to evaluate classification performance in n-fold cross-validation setup.

1. *Micro-Average.* This will count all true positives, false positives, true negatives and false negatives in n folds. From that, precision, recall and F_1 will be compute.

$$Pr_c^\mu = \frac{\sum_{i=1}^{n} |TP_c^{(i)}|}{\sum_{i=1}^{n} \left(|TP_c^{(i)}| + |FP_c^{(i)}| \right)}$$

$$Re_c^\mu = \frac{\sum_{i=1}^{n} |TP_c^{(i)}|}{\sum_{i=1}^{n} \left(|TP_c^{(i)}| + |FN_c^{(i)}| \right)}$$

$$F_{1c}^\mu = \frac{2 Pr_c^\mu Re_c^\mu}{Pr_c^\mu + Re_c^\mu} = \frac{2 \sum_{i=1}^{n} |TP_c^{(i)}|}{\sum_{i=1}^{n} \left(2|TP_c^{(i)}| + |FP_c^{(i)}| + |FN_c^{(i)}| \right)}$$

2. *Macro-Average.* This will average metrics in different folds.

$$Pr_c^M = \frac{\sum_{i=1}^{n} Pr_{1c}^{(i)}}{n}$$

$$Re_c^M = \frac{\sum_{i=1}^{n} Re_{1c}^{(i)}}{n}$$

$$F_{1c}^M = \frac{2 Pr_c^M Re_c^M}{Pr_c^M + Re_c^M}$$

3.5.2 Evaluation on Labeled Users

We use WEKA [14], a machine learning software for data mining tasks. Weka includes a wrapper of libSVM, an implementation of SVM by Chang and Lin [9, 7, 10]. Using 5-fold cross validation, we obtain the performance metrics of the classifiers. Our performance metrics are the standard micro-averaged and macro-averaged precision, recall and F_1 scores for Christian, Muslim, and Buddhist classes. We set the number of top followees used in Social features $k = 100$. In the first set of experiments, we evaluate the performance of our prediction method against the 1050 ground truth labeled users.

Experiment Result

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 additional structural and aggregated features. As shown in Table 3.11, the best macro F_1 scores SVM achieved for Christian, Muslim, and Buddhist classes are 0.891, 0.875, and 0.677 respectively. As shown in Table 3.10, the best macro F_1 scores Naive Bayes achieved for Christian, Muslim, and Buddhist classes are 0.867, 0.842, and 0.644 respectively. Using macro F_1 as the metrics of comparison, we can see that SVM performs better than Naive Bayes by 2.8% for Christian, 3.9% for Muslim, and 5.1% for Buddhist.

Table 3.10: Naive Bayes Performance Macro Precision/Recall/F1

		Christian	Muslim	Buddhist
Self	**Name**	0.778 0.733 0.755	0.689 0.772 0.728	0.000 0.000 0.000
	Content	0.775 0.846 0.809	0.821 0.732 0.774	0.000 0.000 0.000
	Structure	0.605 0.923 0.731	0.895 0.353 0.506	0.000 0.000 0.000
	Aggregated	0.826 0.865 0.845	0.842 0.819 0.830	1.000 0.459 0.629
Social	**Content**	0.801 0.873 0.835	0.846 0.774 0.808	0.000 0.000 0.000
	Structure	0.719 0.926 0.809	0.911 0.569 0.700	1.000 0.265 0.419
	Aggregated	0.805 0.932 0.864	0.930 0.764 0.839	1.000 0.471 0.640
Name + Self + Social		0.819 0.921 0.867	0.899 0.792 0.842	1.000 0.475 0.644

As shown in Table 3.13, the best micro-averaged F_1 scores SVM achieved

Table 3.11: Linear SVM Performance Macro Precision/Recall/F1

			Christian	Muslim	Buddhist
		Name	0.801 0.754 0.777	0.711 0.794 0.750	0.000 0.000 0.000
		Content	0.796 0.867 0.830	0.842 0.753 0.795	0.131 0.044 0.066
Self		**Structure**	0.627 0.945 0.754	0.917 0.374 0.531	0.000 0.000 0.000
		Aggregated	0.848 0.887 0.867	0.863 0.841 0.852	1.000 0.482 0.650
		Content	0.822 0.895 0.857	0.869 0.796 0.831	0.084 0.024 0.037
Social		**Structure**	0.742 0.947 0.820	0.932 0.591 0.723	1.000 0.267 0.421
		Aggregated	0.827 0.953 0.887	0.951 0.786 0.861	1.000 0.496 0.663
Name + Self + Social			0.844 0.943 0.891	0.946 0.815 0.875	1.000 0.512 0.677

for Christian, Muslim, and Buddhist classes are 0.933, 0.912, and 0.696 respectively. As shown in Table 3.12, the best micro-averaged F_1 scores Naive Bayes achieved for Christian, Muslim, and Buddhist classes are 0.911, 0.887, and 0.665 respectively. Using micro-averaged F_1 as the metrics of comparison, we can see that SVM performs better than Naive Bayes by 2.4% for Christian, 2.8% for Muslim, and 4.7% for Buddhist.

Table 3.12: Naive Bayes Performance Micro Precision/Recall/F1

			Christian	Muslim	Buddhist
		Name	0.819 0.774 0.796	0.728 0.812 0.768	0.000 0.000 0.000
		Content	0.815 0.887 0.849	0.851 0.773 0.810	0.000 0.000 0.000
Self		**Structure**	0.648 0.965 0.775	0.937 0.393 0.554	0.000 0.000 0.000
		Aggregated	0.867 0.906 0.886	0.881 0.859 0.870	1.000 0.492 0.657
		Content	0.841 0.914 0.876	0.887 0.814 0.848	0.000 0.000 0.000
Social		**Structure**	0.741 0.967 0.839	0.970 0.630 0.764	1.000 0.308 0.474
		Aggregated	0.846 0.974 0.905	0.970 0.806 0.880	1.000 0.501 0.665
Name + Self + Social			0.864 0.963 0.911	0.952 0.830 0.887	1.000 0.504 0.670

Table 3.13: Linear SVM Performance Micro Precision/Recall/F1

			Christian	Muslim	Buddhist
		Name	0.840 0.795 0.817	0.751 0.835 0.791	0.000 0.000 0.000
		Content	0.837 0.909 0.871	0.873 0.795 0.832	0.182 0.095 0.125
Self		**Structure**	0.669 0.986 0.797	0.959 0.415 0.579	0.000 0.000 0.000
		Aggregated	0.889 0.928 0.908	0.904 0.882 0.893	1.000 0.516 0.681
		Content	0.863 0.936 0.898	0.910 0.837 0.872	0.125 0.048 0.069
Social		**Structure**	0.763 0.988 0.861	0.973 0.632 0.766	1.000 0.333 0.500
		Aggregated	0.869 0.995 0.928	0.992 0.828 0.903	1.000 0.524 0.688
Name + Self + Social			0.886 0.986 0.933	0.977 0.855 0.912	1.000 0.534 0.696

The results in both Tables 3.13, 3.11, 3.12, and 3.10 show that combining all categories of features indeed improves classification performance than using

just one category of features.

Choice of Threshold θ

As shown in Table 3.14, choosing optimal thresholds improves Christian's micro-averaged F_1 from 0.933 to 0.939 (0.6% improvement) when $\theta_{Christian} = 0.057$, Muslim's micro-averaged F_1 from 0.912 to 0.926 (1.5% improvement) when $\theta_{Muslim} = -0.031$, Buddhist's micro-averaged F_1 from 0.696 to 0.722 (3.7% improvement) when $\theta_{Buddhist} = -0.632$.

Table 3.14: Linear SVM Performance with/without Optimal Threshold

	With Default Threshold			With Optimal Threshold		
	Christian ($\theta = 0$)	Muslim ($\theta = 0$)	Buddhist ($\theta = 0$)	Christian ($\theta = 0.057$)	Muslim ($\theta = -0.031$)	Buddhist ($\theta = -0.632$)
F_1^M	0.891	0.875	0.677	0.897	0.889	0.693
F_1^μ	0.933	0.912	0.696	0.939	0.926	0.722

Figure 3.5 shows that optimal threshold give the best F_1, increasing or decreasing thresholds will reduce F_1. In the case of Buddhist class, the F_1 value is very sensitive to the optimal threshold value.

Feature Ranking

As SVM with linear kernel outputs feature weights, Table 3.15 shows top 20 features for the three classes which show the importance of both self and social features with the abbreviations of the feature categories given in Table 3.8. We can see the most important features are the number and fraction of religion tweets in neighborhood tweets, Social-Structural features (top followees), Social and Self-Content features.

Tables 3.16, 3.17, 3.19, and 3.18 give further feature ranking of Self-Name features, Self-Content features, Social-Content features, and Social-Structural features. The tables describe the difference in name n-grams, word usage, and social figures (celebrities, churches, religious websites) of different religions in Singapore. Comparing between different feature types, Social features are
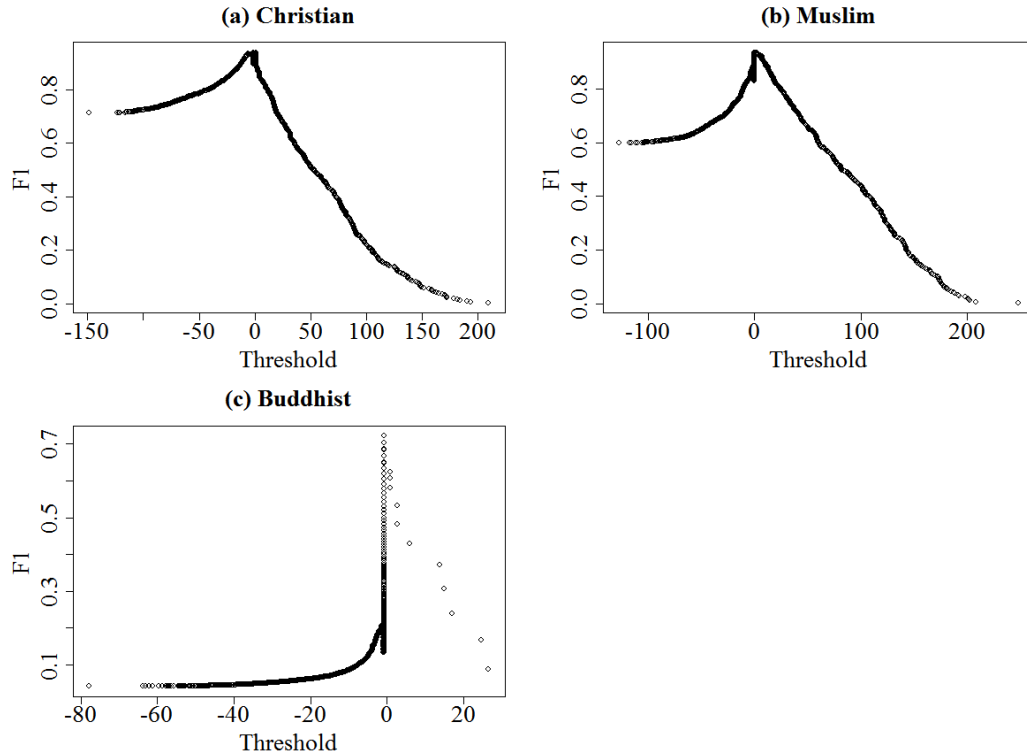
Figure 3.5: Optimal Thresholds for (a) Christian, (b) Muslim, and (c) Buddhist Label

Table 3.15: Top 20 Features By SVM Weight

Rank	Christian	Muslim	Buddhist
1	SO-AG-NRT(Christian) (Number of Christian Tweets)	SO-ST (IslamSpeaks)	SO-ST (AjahnBrahm)
2	SO-AG-NFT(Christian) (Fraction of Christian Tweets)	SO-ST (IslamicThinking)	SO-AG-NRT(Buddhist) (Number of Buddhist Tweets)
3	SO-CO-NH-BI (jesus)	SE-CO-BI (allah)	SO-AG-NFT(Buddhist) (Fraction of Buddhist Tweets)
4	SO-CO-NH-TI (jesus)	SE-CO-TI (allah)	SO-ST (tinybuddha)
5	SE-AG-AMID(Christian) (Christian Indeg Is Max)	SO-AG-NRT(Muslim) (Number of Muslim Tweets)	SO-ST (Zen_Moments)
6	SE-CO-BI (bible)	SO-AG-FRT(Muslim) (Fraction of Muslim Tweets)	SE-CO-BI (meditation)
7	SE-CO-TI (bible)	SO-CO-NH-BI (quran)	SE-CO-TI (meditation)
8	SO-ST (SgCatholic)	SO–CO-NH-TI (quran)	SO-ST (thedailybeast)
9	SO-CO-NH-BI (pastor)	SO-CO-NH-BI (maghrib)	SO-ST (TheMomsView)
10	SO-CO-NH-TI (pastor)	SO-CO-NH-TI (maghrib)	SO-ST (Buddhism_Now)
11	SO-CO-NH-BI (testimony)	SO-CO-NH-BI (jannah)	SO-ST (elephantjournal)
12	SO-CO-NH-TI (testimony)	SO-CO-NH-TI (jannah)	SE-CO-BI (rinpoche)
13	SO-ST (JoyceMeyer)	SE-CO-BI (iftar)	SE-CO-BI (rinpoche)
14	SO-CO-NH-BI (prov)	SE-CO-TI (iftar)	SE-CO-NH-BI (rinpoche)
15	SO-CO-NH-TI (prov)	SO-ST (TheNobleQuran)	SO-CO-NH-TI (rinpoche)
16	SO-CO-NH-BI (svc)	SO-CO-NH-BI (tgk)	SE-CO-BI (periodically)
17	SO-CO-NH-TI (svc)	SO-CO-NH-TI (tgk)	SE-CO-TI (periodically)
18	SO-CO-NH-BI (romans)	SO-CO-NH-BI (imam)	SO-CO-NH-BI (periodically)
19	SO-CO-NH-TI (romans)	SO-CO-NH-TI (imam)	SO-CO-NH-TI (periodically)
20	SO-ST (hillsongunited)	SO-CO-NH-BI (kene)	SE-CO-BI (openness)

more important than Self features, Structural features are more important than Content features and Name features.

Christian names usually include "hua" *Joshua Heng*, *Melinda Chua*, and

Jennifer Z<u>huang</u>, as well as "go" *Sean <u>Goh</u>*, *Gre<u>go</u>ry Sim*, and Diego Gonzalez R..
Muslim names usually include "sya" *Nadiah <u>Sya</u>zwani*, *<u>Sya</u>fa'at Salleh*, as well
as "ila" like *Khal<u>ila</u>*, *Faz<u>ila</u>i*. Buddhist names usually include "drol" *Nam<u>drol</u>*
Donyo, *Chonyimin<u>drol</u>*.

Table 3.16: Top 20 SE-NA($\langle ngram \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	hua (256)	sya (155)	drol (143)
2	go (257)	ila (157)	dro (145)
3	gr (260)	nur (160)	orth (146)
4	leo (262)	af (162)	ishn (148)
5	hoo (265)	uha (166)	iro (154)
6	nne (271)	rul (168)	ndrol (159)
7	lvi (276)	muh (175)	hnan (167)
8	koh (284)	hamma (181)	kms (175)
9	oe (292)	hamm (187)	yimin (184)
10	enn (301)	amma (195)	mind (194)
11	avi (311)	mmad (204)	kmsp (205)
12	jos (323)	amm (214)	onyi (217)
13	gra (336)	ammad (225)	onyo (230)
14	iel (350)	muha (238)	ishna (244)
15	eli (365)	muham (240)	hii (259)
16	sea (381)	uham (254)	gto (275)
17	lvin (398)	uhamm (269)	lnes (292)
18	eong (416)	uru (285)	gt (310)
19	nic (435)	urul (302)	dony (329)
20	vi (455)	irah (321)	pg (349)

Important words to predict a user as Christian are "bible", "jesus", "church" while important words to predict a user as Muslim are "allah", "iftar", "masjid", "quran", important words to predict a user as Buddhist are "meditation", "openess", "mindfulness". These words are indeed religion related words.

Important religion words from neighborhood to predict a user as Christian are "jesus", "pastor", "testimony" while religion words from neighborhood to predict a user as Muslim are "quran", "maghrib", "jannah", religion words from neighborhood to predict a user as Buddhist are "rinpoche", "periodically", "openness".

The important followees that Christians follow are *SgCatholic* (Singapore Catholics), *JoyceMeyer*, *hillsongunited* while followees that Muslims follow are

Table 3.17: Top 20 SE-CO-BI($\langle word \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	bible (6)	allah (3)	meditation (6)
2	jesus (33)	iftar (13)	rinpoche (12)
3	church (94)	masjid (45)	periodically (16)
4	testimony (107)	quran (49)	openness (20)
5	pastor (138)	awak (53)	mindfulness (24)
6	christ (143)	maghrib (69)	composer (26)
7	hillsong (176)	kene (73)	tomtom (30)
8	psalm (180)	pakai (89)	prequel (41)
9	revival (198)	alhamdulillah (91)	visualization (43)
10	nak (201)	pasal (99)	defies (47)
11	chc (203)	abeh (105)	corporations (52)
12	fellowship (218)	terawih (114)	hinduism (57)
13	gospel (225)	korang (116)	konjam (58)
14	anointing (230)	cakap (119)	mouthful (71)
15	cornerstone (233)	takpe (128)	une (79)
16	corinthians (241)	ade (133)	bwahahahaha (84)
17	amen (244)	ckp (136)	indifferent (85)
18	hogc (248)	insyaallah (142)	dharma (92)
19	proverbs (251)	tengok (148)	emperor (97)
20	churches (254)	takde (153)	dalai (104)

Table 3.18: Top 20 SO-CO-NH-BI($\langle word \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	jesus (3)	quran (7)	rinpoche (13)
2	pastor (9)	maghrib (9)	periodically (17)
3	testimony (11)	jannah (11)	openness (21)
4	prov (14)	tgk (16)	composer (27)
5	svc (16)	imam (18)	tomtom (31)
6	romans (18)	kene (20)	aaaaahh (38)
7	prosperous (21)	pulak (22)	visualization (44)
8	christ (23)	kerana (24)	defies (48)
9	anointing (28)	madrasah (28)	corporations (53)
10	psalm (35)	ade (30)	hinduism (59)
11	dismayed (37)	insyallah (33)	konjam (60)
12	revival (43)	ustaz (35)	mouthful (72)
13	chc (45)	hijab (37)	une (80)
14	churchill (47)	insya (39)	bwahahahaha (86)
15	hillsong (52)	azan (41)	indifferent (87)
16	teresa (57)	rasulullah (42)	dharma (93)
17	harvest (61)	bukhari (47)	emperor (98)
18	bernard (66)	darul (51)	dalai (105)
19	forsaking (67)	sentiasa (55)	abrsm (115)
20	pringle (68)	jom (57)	buddhism (118)

IslamSpeaks, IslamicThinking, TheNobleQuran, important followees that Buddhists follow are *AjahnBrahm, tinybuddha, Zen_Moments*. These users are popular religion organizations, pastors, or authors.

Table 3.19: Top 20 SO-ST($\langle followee \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	SgCatholic (8)	IslamSpeaks (1)	AjahnBrahm (1)
2	JoyceMeyer (13)	IslamicThinking (2)	tinybuddha (4)
3	hillsongunited (20)	TheNobleQuran (15)	Zen_Moments (5)
4	RickWarren (25)	banchothematrep (26)	thedailybeast (8)
5	DarleneZschech (26)	MaherZain (27)	TheMomsView (9)
6	JohnPiper (30)	TaufikBatisah (32)	Buddhism_Now (10)
7	Seowhow (31)	LisaSurihani (59)	elephantjournal (11)
8	JohnBevere (32)	IslamQuotes (62)	bobdylan (34)
9	TGC (39)	quran (75)	daily_buddhism (35)
10	JLin7 (40)	Shaheizy_Sam (76)	theworsthorse (36)
11	HarrisJosh (41)	Hadithoftheday (77)	Bodhipaksa (37)
12	ARBernard (42)	WardinaSafiyyah (80)	VincentHorn (40)
13	ReinhardBonnke (49)	Aaron535Aziz (81)	waylonlewis (51)
14	Lia_Chan (50)	iloveAllaah (86)	lessig (56)
15	hillsong (51)	IslamicMoments (87)	Swamy39 (65)
16	philpringle (54)	islamicthought (88)	the_hindu (66)
17	desiringgod (55)	ErraFaziraWC (93)	DhammaLinks (67)
18	mynameissun (56)	MuslimMoments (94)	BuddhistGeeks (68)
19	SidMohede (59)	DailyHadiths (97)	DharmaDots (69)
20	MaxLucado (60)	DailyHadith (98)	tricyclemag (70)

If we look further into these followees shown in Tables 3.20, 3.21, and 3.22, we can see that the majority of Christian top followees are Christian celebrities and churches. The majority of Muslim and Buddhist top followees, however, are general religion Twitter accounts.

3.5.3 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 so we rank them by SVM output scores. We then manually judged the top 100 users under the Christian class, top 100 users under the Muslim class, and top 100 users under the Buddhist class. Table 3.23 shows the top

Table 3.20: Top 20 Christian SO-ST($\langle followee \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Description
1	SgCatholic (8)	Singapore Catholics
2	JoyceMeyer (13)	Christian book writer
3	hillsongunited (20)	Christian website
4	RickWarren (25)	Christian pastor, book writer
5	DarleneZschech (26)	Christian singer
6	JohnPiper (30)	Christian pastor
7	Seowhow (31)	Heart of God Church pastor
8	JohnBevere (32)	Christian pastor
9	TGC (39)	Christian website
10	JLin7 (40)	Christian celebrity
11	HarrisJosh (41)	Christian pastor, book writer
12	ARBernard (42)	Christian CEO
13	ReinhardBonnke (49)	Christian celebrity
14	Lia_Chan (50)	Christian celebrity
15	hillsong (51)	Hillsong church
16	philpringle (54)	Christian celebrity
17	desiringgod (55)	Christian pastor
18	mynameissun (56)	Christian celebrity
19	SidMohede (59)	Christian celebrity
20	MaxLucado (60)	Christian celebrity

Table 3.21: Top 20 Muslim SO-ST($\langle followee \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Muslim	Description
1	IslamSpeaks (1)	Islamic Twitter account
2	IslamicThinking (2)	Islamic Twitter account
3	TheNobleQuran (15)	Islamic Twitter account
4	banchothematrep (26)	Muslim celebrity
5	MaherZain (27)	Muslim celebrity
6	TaufikBatisah (32)	Muslim celebrity
7	LisaSurihani (59)	Muslim celebrity
8	IslamQuotes (62)	Islamic Twitter account
9	quran (75)	Islamic Twitter account
10	Shaheizy_Sam (76)	Muslim celebrity
11	Hadithoftheday (77)	Islamic Twitter account
12	WardinaSafiyyah (80)	Muslim celebrity
13	Aaron535Aziz (81)	Muslim celebrity
14	iloveAllaah (86)	Islamic Twitter account
15	IslamicMoments (87)	Islamic Twitter account
16	islamicthought (88)	Islamic Twitter account
17	ErraFaziraWC (93)	Muslim celebrity
18	MuslimMoments (94)	Islamic Twitter account
19	DailyHadiths (97)	Islamic Twitter account
20	DailyHadith (98)	Islamic Twitter account

Table 3.22: Top 20 Buddhist SO-ST($\langle followee \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Buddhist	Description
1	AjahnBrahm (1)	Monk
2	tinybuddha (4)	Buddhist Twitter account
3	Zen_Moments (5)	Buddhist Twitter account
4	thedailybeast (8)	News Twitter account
5	TheMomsView (9)	News Twitter account
6	Buddhism_Now (10)	Buddhist Twitter account
7	elephantjournal (11)	Buddhist Twitter account
8	bobdylan (34)	Singer
9	daily_buddhism (35)	Buddhist Twitter account
10	theworsthorse (36)	Buddhist Twitter account
11	Bodhipaksa (37)	Buddhist celebrity
12	VincentHorn (40)	Buddhist celebrity
13	waylonlewis (51)	Buddhist celebrity
14	lessig (56)	Twitter celebrity
15	Swamy39 (65)	Twitter celebrity
16	the_hindu (66)	India's National Newspaper
17	DhammaLinks (67)	Buddhist Twitter account
18	BuddhistGeeks (68)	Buddhist Twitter account
19	DharmaDots (69)	Buddhist Twitter account
20	tricyclemag (70)	Buddhist Twitter account

20 users ranked by SVM scores. Our experiments have shown that members of different religions differ in the ways they tweet and follow. Therefore, the manual judgement was conducted by (1) checking their profile pages including biography; (2) examining religion related tweets generated by the users; and (3) checking his religion related followees and followers. The results show that the top users under Christian and Muslim classes are indeed assigned with the correct religions,i.e., the precision of the manual check is 100%. There are 35 Buddhists correctly predicted. For example, the top user under the Buddhist class, *scantm*, said the following in his biography:

> "Educator. L&T Technologist. Open Source Champion.Community Builder. *Dharma Student.* Astronomer. Dive Master. Fringe"

Another top user under the Buddhist class, *KwanYinChanLin*, is the twitter account of Kwan Um School of Zen, Singapore. The top user under the Christian class, *Calvin_Lee_*, tweeted about his religion:

"*Church*, family and studies. My tripod, my responsibility. I will uphold them well"

Another top user under Christian class, *chewdawei*, tweeted about his church service:

"Getting ready for service at Hope *Church* Brisbane in UQ! Woohoo!".

The top user under Muslim class, *ManutdZul*, tweeted in a Muslim way:

"my future marriage shall have this, *insya'allah* (; hehehe :p",

Another top user under Muslim class, *syaaafiqah*, tweeted that

"@Svuee You look so down.Babe, *Allah* knws best. Whatever happens, just tawakal and redha coz you knw you tried your best. HE will do the rest".

Table 3.23: Top 20 Users By SVM Score

Rank	Christian	Muslim	Buddhist
1	Calvin_Lee_	ManutdZul	seantm
2	chewdawei	syaaafiqah	KwanYinChanLin
3	zeewhy	Nisa07021992	SalivaVagaries
4	jamieleesj	nrfkhrh	namdrol_Dekyi
5	johntann	AndreRoslan	Nisha_L
6	j_fen	LiZiLiCiouZ	postmuseum
7	yassychan	superbanso	twittfrog
8	puahsihui	syirah_nasyitah	alightheart
9	konghee	Edriee	greenteacup
10	jemquek	SuzzySues	4nirav
11	KianLeng	CAca_dong	bruceshou
12	JianMingTan	aydaisnin	GeraldineNord
13	Chris_Honegger	nurulsuperduper	_wakeupnow_
14	garrettleejw	Ukhti_Bilah	k_rohit_a
15	deckstor	brokenyeul	Sgthinker
16	Glenn_Yong	Batrishiaa	kimballchilli
17	wilsonbarnabas	ShikinMajid	milkpudding04
18	jansontow	NoninieyyDrew	potatofantasy
19	Ngbingrong	zoolayCAR	eYeka
20	Lawrence_LeeCS	derpetteeee	charrleneee

Table 3.24: Top 20 Christians By SVM Scores

Rank	Christian	Description
1	Calvin_Lee_	Member of Heart of God Church
2	chewdawei	Member of Heart of God Church
3	zeewhy	Member of Heart of God Church
4	jamieleesj	Member of Heart of God Church
5	johntann	Member of Heart of God Church
6	j_fen	Member of Heart of God Church
7	yassychan	Member of Heart of God Church
8	puahsihui	Member of Heart of God Church
9	konghee	Christian pastor
10	jemquek	Member of Heart of God Church
11	KianLeng	Member of Heart of God Church
12	JianMingTan	Christian celebrity
13	Chris_Honegger	Member of Heart of God Church
14	garrettleejw	Heart of God Church pastor
15	deckstor	Member of Heart of God Church
16	Glenn_Yong	Member of Heart of God Church
17	wilsonbarnabas	Member of City Harvest Church
18	jansontow	Member of Heart of God Church
19	Ngbingrong	Member of Heart of God Church
20	Lawrence_LeeCS	Member of Heart of God Church

3.6 Conclusion

In this chapter, we have shown the benefits of including a various types of features for religion prediction. The classifiers trained using the whole set of features perform better than the ones trained using just subsets of features. By choosing SVM's optimal threshold θ, we are able to improve F_1 for all 3 classes over those using the default threshold $\theta = 0$. Social features are more important than Self features in revealing user's religion. Also, structural features are more important than content features and name features. We also apply the method on all users without labels. The precision of top 100 users for the three classes are 100%, 100% and 35% for the Christian, Muslim and Buddhlist classes respectively. In the next Chapter, we will show how this accuracy can be further improved by introducing collective classification.

Table 3.25: Top 20 Muslims By SVM Scores

Rank	Muslim	Description
1	ManutdZul	Muslim, full name: Muhammad Zulhusni
2	syaaafiqah	Muslim, full name: Nur Syafiqah Z.
3	Nisa07021992	Muslim, full name: Khairunnisa razif
4	nrfkhrh	Muslim, full name: Kyra Nurfakhirah
5	AndreRoslan	Muslim, said "Insya-Allah" many times
6	LiZiLiCiouZ	Muslim, full name: Liz Suriyani
7	superbanso	Muslim, full name: farahdiyanah mohdrazid
8	syirah_nasyitah	Muslim, full name: Basyirah ASHBURN
9	Edriee	Muslim, full name: Muhammad Edrie Rizwan
10	SuzzySues	Muslim, said "Allah" many times
11	CAca_dong	Muslim, said "Allahuakbar" many times
12	aydaisnin	Muslim, said "Allah" many times
13	nurulsuperduper	Muslim, full name: Nurul Ain
14	Ukhti_Bilah	Muslim, said "Allah" many times
15	brokenyeul	Muslim, said "I thank Allah for my wonderful parents."
16	Batrishiaa	Muslim, full name: Nadhrah Batrishia
17	ShikinMajid	Muslim, full name: Nurashikin Majid
18	NoninieyyDrew	Muslim, tweeted "So soak in the moment, use it well, and remember where it started. Allah is great."
19	zoolayCAR	Muslim, tweeted "May allah bless you with good health and wealth."
20	derpetteeee	Muslim, tweeted "Islam is my subject, Quran is my textbook, Prophet Muhammad SAW is my teacher, Dunya is my test, Allah is my Judge."

Table 3.26: Top 20 Buddhists By SVM Scores

Rank	Buddhist	Description
1	seantm	Buddhist, Dharma Student, Follower of Zen Moments, Buddhist Geeks, and Buddhism Now
2	KwanYinChanLin	Kwan Um School of Zen, Singapore, Follower of Zen Moments, Buddhist Geeks, and Buddhism Now
3	SalivaVagaries	Buddhist, Follower of Zen Moments
4	namdrol_Dekyi	Buddhist, Follower of Buddha Quotes
5	Nisha_L	Buddhist, Follower of Tiny Buddha
6	postmuseum	Buddhist, Follower of Elephant Journal, Tricycle Magazine, and Thich Nhat Hanh
7	twittfrog	Buddhist, Follower of Elephant Journal
8	alightheart	Buddhist, Follower of Bodhipaksa and Elephant Journal
9	greenteacup	Buddhist, Follower of Smart Buddhist
10	4nirav	Not confirmed
11	bruceshou	Not confirmed
12	GeraldineNord	Not confirmed
13	_wakeupnow_	Buddhist, tweeted "buddhism, science, symposium iv growing a beautiful mind"
14	k_rohit_a	Not confirmed
15	Sgthinker	Buddhist, Follower of Thich Nhat Hanh, Daily Buddhism, and Buddhism Now
16	kimballchilli	Not confirmed
17	milkpudding04	Not confirmed
18	potatofantasy	Not confirmed
19	eYeka	Not confirmed
20	charrleneee	Not confirmed

47

Chapter 4

Collective Religion Prediction

In this section, we give a detailed description of another method, Collective Classification, and the experiments that evaluate the method.

4.1 Collective Classification Method

In collective classification, we still have three classifiers, one for each religion. In our problem context, there are very few labeled users. The key idea of collective classification is to use a few predicted user labeled as additional labeled data. The method performs classification in multiple iterations. In each iteration, we add the top scored users of each religion label to the labeled data before recomputing label-dependent features so as to retrain the classifiers. The users predicted with highest confidence scores are obviously the good choices. If these users are correctly predicted, collective classification will be able to learn better and produce more accurate prediction [23, 30, 12, 21].

We present the steps to derive label-dependent features. We then iteratively derive features, train classifiers and apply to unlabeled users.

4.1.1 Label-Dependent Features

Our propose collective classification method utilizes both label-independent and label-dependent features. The former is computed only once in the first

iteration of the classification process because label-independent features are static. The label-dependent features are recomputed every iteration because the labeled data are extended to include top scored users of each label. One can refer to Section 3.3 for the computation of label-independent features.

Label-dependent features include Self and Social-Structural features, Social-Content features, and Self and Social-Aggregated features as they are dependent of labeled data. The algorithm to compute label-dependent features is presented in Algorithm 1. Firstly, each user's Self-Structural features (RInDegree, label-indegrees, label-outdegrees, etc.) are computed (denoted as K). Secondly, Self-Aggregated features, (denoted as B), can be computed from users' Self-Content features and Self-Structural features. Thirdly, Social-Content features, is computed by combining Self-Content features (T) from user's neighbors and neighborhood. Fourthly, each user's Social-Structural features (denoted as H) is computed from Self-Structural features. Fifthly, Social-Aggregated features, number of religious tweets in user's top followees, min/ average/ max RInDegree, indegree of user's top followees, is computed from Social-Content feature (S) and Social-Structural features (H). The algorithm is illustrated in Figure 4.1.

Input:
V–set of users,
V_L–set of labeled users,
T–users' self content features
Output:
K–Self-Structural features,
H–Social-Structural features,
B–Self-Aggregated features,
S–Social-Content features,
A–Social-Aggregated features
foreach $user\ u \in V$ **do**
    $K(v) \leftarrow$ ComputeDegreeFeatures(V_L);
    $B(v) \leftarrow$ ComputeSelfAggregatedFeatures($K(v)$, V_L);
    $S(v) \leftarrow$ CombineNeighborhoodFeatures(T, V_L);
    $H(v) \leftarrow$ ComputeTopFolloweesFeatures($K(v)$, V_L);
    $A(v) \leftarrow$ ComputeSocialAgregatedFeatures(S, H, V_L);
end

Algorithm 1: Label-Dependent Feature Derivation
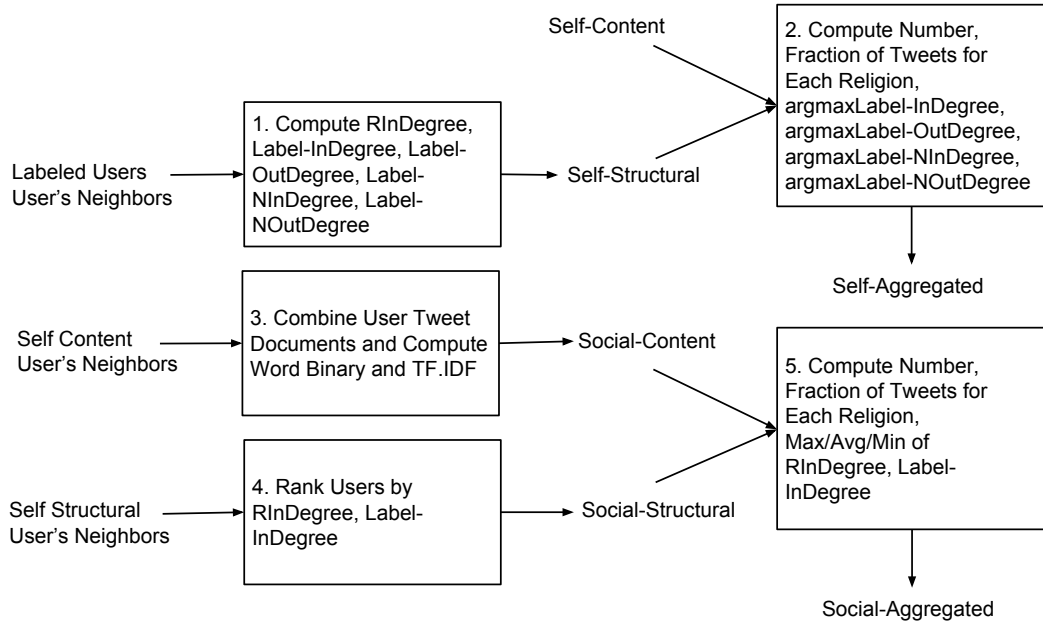
Figure 4.1: Label-Dependent Feature Derivation

4.1.2 Collective Classification Algorithm

The full algorithm is presented in Algorithm 2. In first iteration, label-independent and label-dependent features are derived. For each religion label, a classifier is trained for each religion using the labeled user data. We then apply the classifier on the unlabeled users and derive a confidence score for each of them. The top $p\%$ confidence score users of each religion are included as pseudo-labeled users (V_R') which are added to training data in the next iteration. SVM optimal thresholds θ_R are obtained by choosing the ones that maximize F_1. Naive Bayes does need to choose these thresholds.

In each following iteration, we recompute label-dependent features using both the labeled and pseudo-labeled users. After that, we train a new SVM classifier for each religion label. We label more $p\%$ (of the number of labeled data) top confidence score users. We repeat the process i_{max} times ($i_{max} = 10$ by default).

Input:

V–user set,

V_R–set of labeled users of religion R,

i_{max}–maximum number of iteration,

θ_R–optimal SVM threshold for religion R,

$p\%$–percentage of additional instances increased in each iteration

Output:

C_R–classifier for religion R where R is "Christian", "Muslim", and "Buddhist"

$I \leftarrow \text{ComputeLabelIndependentFeatures}(V)$;

$i \leftarrow 0$;

foreach $R \in \{\text{"}Christian\text{"}, \text{"}Muslim\text{"}, \text{"}Buddhist\text{"}\}$ **do**

$\quad | \quad V_R' \leftarrow \varnothing$

end

while $i \leq i_{max}$ **do**

$\quad | \quad i \leftarrow i + 1$;

$\quad | \quad V_L \leftarrow \bigcup_R (V_R \cup V_R')$;

$\quad | \quad J \leftarrow \text{ComputeLabelDependentFeatures}(V, V_L)$;

$\quad | \quad$ **foreach** $R \in \{\text{"}Christian\text{"}, \text{"}Muslim\text{"}, \text{"}Buddhist\text{"}\}$ **do**

$\quad | \quad \quad | \quad C_R \leftarrow \text{TrainClassifier}(I, J, V_L)$;

$\quad | \quad \quad | \quad Conf_R \leftarrow C_R(I, J, V - \bigcup_R V_R)$;

$\quad | \quad \quad | \quad V_R' \leftarrow \{v | (v \in V - \bigcup_R V_R) \wedge (Conf_R(v) \geq$

$\quad | \quad \quad | \quad \theta_R) \wedge (rank(Conf_R(v)) \leq |V_R| \cdot i \cdot p/100)\}$;

$\quad | \quad$ **end**

end

Algorithm 2: Collective Classification

4.2 Experiments

4.2.1 Comparison with Non-Collective Classification

In this section, we will compare the collective classification method with the non-collective classification method presented in previous section. We run 5-fold cross validation using Naive Bayes and linear SVM with optimal thresholds $\theta_{Christian} = 0.057$, $\theta_{Muslim} = -0.031$, and $\theta_{Buddhist} = -0.632$ which maximize F_1 of SVM on labeled data (5-fold cross validation), percentage of additional pseudo-labeled users in each iteration $p = 10$, and maximum number of iterations $i_{max} = 10$. Both macro-averaged F_1 (F_1^M) and micro-averaged F_1 (F_1^μ) are reported. As shown in Table 4.1, the collective classification method yields about 2 to 5% improvement in F_1^μ and F_1^M for both linear SVM and Naive Bayes. Linear SVM's performances are 2 to 5 % better than Naive Bayes'

performances.

Table 4.1: Non-Collective Classification vs. Collective Classification with $p = 10$, $i_{max} = 10$

	Non-Collective Classification			Collective Classification		
	Christian	Muslim	Buddhist	Christian	Muslim	Buddhist
Naive Bayes F_1^M	0.867	0.842	0.644	0.882	0.862	0.724
Linear SVM F_1^M	0.897	0.889	0.693	0.934	0.932	0.746
Naive Bayes F_1^μ	0.911	0.887	0.670	0.926	0.919	0.705
Linear SVM F_1^μ	0.939	0.926	0.722	0.957	0.951	0.754

4.2.2 Performance in Different Iterations

The performance in different iterations using macro-averaged F_1 (F_1^M) and micro-averaged F_1 (F_1^μ) of Collective Classification (with $p = 10$) when using Naive Bayes or linear SVM are shown in Tables 4.2 and 4.3 respectively. We observe that the performance of all three classes is improved in almost all iterations. Linear SVM performs better than Naive Bayes by 2 to 5%.

Table 4.2: Collective Classification F_1^M (Linear SVM, Naive Bayes) in Multiple Iterations with $p = 10$, $i_{max} = 10$

Iteration	Linear SVM			Naive Bayes		
	Christian	Muslim	Buddhist	Christian	Muslim	Buddhist
0	0.897	0.889	0.693	0.867	0.842	0.644
1	0.903	0.896	0.708	0.875	0.845	0.689
2	0.908	0.902	0.717	0.876	0.848	0.696
3	0.912	0.907	0.722	0.877	0.851	0.701
4	0.916	0.912	0.726	0.877	0.853	0.705
5	0.920	0.917	0.732	0.878	0.855	0.710
6	0.923	0.921	0.736	0.878	0.857	0.713
7	0.926	0.924	0.739	0.881	0.859	0.717
8	0.929	0.928	0.743	0.881	0.859	0.719
9	0.933	0.930	0.746	0.882	0.862	0.722
10	0.934	0.932	0.746	0.882	0.862	0.724

4.3 Experiments with Varying p

In each iteration, we add top p% more highest classification score users of each religion to the already labeled users as pseudo-labeled users to train classifiers.

Table 4.3: Collective Classification F_1^μ (Linear SVM, Naive Bayes) in Multiple Iterations with $p = 10$, $i_{max} = 10$

	Linear SVM			Naive Bayes		
Iteration	Christian	Muslim	Buddhist	Christian	Muslim	Buddhist
0	0.939	0.926	0.722	0.911	0.887	0.670
1	0.942	0.929	0.725	0.914	0.891	0.675
2	0.944	0.933	0.728	0.917	0.907	0.679
3	0.946	0.936	0.731	0.919	0.909	0.684
4	0.948	0.939	0.735	0.921	0.911	0.689
5	0.950	0.942	0.738	0.923	0.913	0.693
6	0.952	0.944	0.742	0.923	0.915	0.697
7	0.954	0.947	0.745	0.924	0.917	0.699
8	0.956	0.949	0.748	0.924	0.917	0.701
9	0.957	0.951	0.752	0.926	0.919	0.703
10	0.957	0.951	0.754	0.926	0.919	0.705

The performance of the collective classification with $p = 5$, 10, 20, 40, and 80 with with $i_{max} = 10$ are shown in Tables 4.4 and 4.5. We observe that $p = 10$ gives better results than $p = 5$. However, increasing p further more does not yield better results. Linear SVM's results are 2 to 5% better than Naive Bayes.

Table 4.4: Collective Classification F_1^M (Linear SVM, Naive Bayes) with Varying p, $i_{max} = 10$

	Linear SVM			Naive Bayes		
p	Christian	Muslim	Buddhist	Christian	Muslim	Buddhist
5	0.920	0.917	0.732	0.878	0.855	0.710
10	0.934	0.932	0.746	0.882	0.862	0.724
20	0.932	0.930	0.741	0.881	0.860	0.720
40	0.927	0.925	0.738	0.879	0.856	0.713
80	0.919	0.918	0.732	0.876	0.843	0.702

Table 4.5: Collective Classification F_1^μ (Linear SVM, Naive Bayes) with Varying p, $i_{max} = 10$

	Linear SVM			Naive Bayes		
p	Christian	Muslim	Buddhist	Christian	Muslim	Buddhist
5	0.950	0.942	0.738	0.923	0.913	0.693
10	0.957	0.951	0.754	0.926	0.919	0.705
20	0.957	0.951	0.754	0.925	0.917	0.701
40	0.955	0.948	0.746	0.924	0.915	0.697
80	0.948	0.944	0.738	0.921	0.914	0.692

4.4 Feature Ranking

From labeled and pseudo-labeled data, we apply our collective classification method using a linear SVM with $p = 10$ and examine the feature importance. As the collective classification method requires multiple iterations, we consider the feature ranking generated by the final iteration of the method shown in Table 4.6. Table 4.6 shows that among top 20 features, 16 are label-dependent features. Therefore, we conclude that label-dependent features are more important than label-independent features. Among label-dependent features, aggregated features are more important than structural features and content features.

Table 4.6: Top 20 Features By SVM Weight

Rank	Christian	Muslim	Buddhist
1	SO-AG (Number of Christian Tweets)	SO-AG (Number of Muslim Tweets)	SO-AG (Number of Buddhist Tweets)
2	SO-AG (Fraction of Christian Tweets)	SO-AG (Fraction of Muslim Tweets)	SO-AG (Fraction of Buddhist Tweets)
3	SE-AG (Number of Christian Tweets)	SE-AG (Number of Muslim Tweets)	SE-AG (Number of Buddhist Tweets)
4	SE-AG (Fraction of Christian Tweets)	SE-AG (Fraction of Muslim Tweets)	SE-AG (Fraction of Buddhist Tweets)
5	SO-AG (Max Christian label-indeg)	SO-AG (Max Muslim label-indeg)	SO-ST (Zen_Moments)
6	SO-AG (Average Christian label-indeg)	SO-AG (Average Muslim label-indeg)	SO-ST (Buddhism_Now)
7	SE-AG (Christian label-indeg is max)	SE-AG (Muslim label-indeg is max)	SO-ST (elephantjournal)
8	SE-AG (Christian label-Nindeg is max)	SE-AG (Muslim label-Nindeg is max)	SO-ST (DhammaLinks)
9	SO-ST (konghee)	SO-ST (IslamicThinking)	SO-ST (theworsthorse)
10	SO-ST (JoyceMeyer)	SO-ST (IslamSpeaks)	SE-ST (Buddhist label-indeg)
11	SE-ST (Christian label-indeg)	SE-ST (Muslim label-indeg)	SE-ST (Buddhist label-Nindeg)
12	SE-ST (Christian label-Nindeg)	SE-ST (Muslim label-Nindeg)	SO-CO-NH-BI (mindfulness)
13	SO-CO-NH-BI (pastor)	SO-CO-NH-BI (insyaallah)	SO-CO-NH-TI (mindfulness)
14	SO-CO-NH-TI (pastor)	SO-CO-NH-TI (insyaallah)	SO-CO-NH-BI (meditation)
15	SO-CO-NH-BI (church)	SO-CO-NH-BI (masjid)	SO-CO-NH-TI (meditation)
16	SO-CO-NH-TI (church)	SO-CO-NH-TI (masjid)	SO-CO-NH-BI (dalai)
17	SE-CO-BI (psalm)	SE-CO-BI (allah)	SE-CO-TI (dalai)
18	SE-CO-TI (psalm)	SE-CO-TI (allah)	SE-CO-BI (dharma)
19	SE-CO-BI (jesus)	SE-CO-BI (insyaallah)	SE-CO-TI (dharma)
20	SE-CO-TI (jesus)	SE-CO-TI (insyaallah)	SE-CO-BI (indifferent)

The feature names are explained in Table 3.8. Tables 4.7, 4.8, 4.10, and 4.9 give feature ranking of Self-Name, Self-Content, Social-Content, and Social-Structural features respectively. As shown in Table 4.7, Christian names usually include "leo" *Jared* <u>*Leo*</u>, <u>*Cleo*</u>., and <u>*Leo*</u>*nard Jonathan Oh*, or "gra"

Grace Choong, *Gracema Lo*, and *Davidkingraj*. Muslim names usually include "nur" *Nurul Syazwanie*, *nur ameesha*, and *Nur Insyirah*, or "hamm" *MuhammadRusydi-Rosli*, *Mohammad Faris*, and *Muhammad Shariff*. Buddhist names usually include "mind" *Chonyimindrol* and *Mindfulness*.

Table 4.7: Top 20 SE-NA($\langle ngram \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	leo (144)	nur (151)	mind (143)
2	gra (149)	hamm (154)	hnan (145)
3	eong (157)	sya (161)	kmsp (146)
4	hua (162)	hamma (163)	onyi (148)
5	hoo (165)	ila (166)	iro (154)
6	vi (169)	rul (171)	onyo (159)
7	lvi (172)	muh (174)	dro (167)
8	go (177)	af (179)	kms (175)
9	oe (180)	uha (182)	yimin (184)
10	enn (189)	amma (187)	orth (194)
11	avi (192)	mmad (191)	drol (205)
12	jos (196)	amm (196)	ishn (217)
13	koh (199)	urul (198)	ndrol (230)
14	iel (202)	muha (202)	ishna (244)
15	eli (204)	uru (208)	pg (259)
16	sea (207)	uham (211)	gto (275)
17	lvin (211)	irah (216)	dony (292)
18	gr (214)	muham (224)	gt (210)
19	nic (221)	ammad (229)	lnes (229)
20	nne (233)	uhamm (231)	hii (249)

Table 4.8 shows that important words to predict a user as Christian are "psalm", "jesus", and "church". Important words to predict a user as Muslim are "allah", "insyaallah", "quran", and "quran". Important words to predict a user as Buddhist are "mindfulness", "dharma", and "openness". Those are typical words of these religions.

As shown in Table 4.9, the important social words to predict a user as Christian are "pastor", "church", and "hillsong" while the important social words to predict a user as Muslim are "insyaallah", "masjid", and "terawih". For Buddhist, the important social words to predict a user as Buddhist are "mindfulness", "meditation", and "dalai" as shown in Table 4.9.

As shown in Table 4.10, the important followees that Christians follow are

Table 4.8: Top 20 SE-CO-NH-BI($\langle word \rangle$)s By SVM Weights (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	psalm (17)	allah (17)	dharma (18)
2	jesus (19)	insyaallah (19)	indifferent (20)
3	church (49)	quran (34)	openness (38)
4	testimony (57)	masjid (43)	meditation (42)
5	pastor (59)	awak (45)	dalai (55)
6	christ (69)	maghrib (50)	buddhism (57)
7	hillsong (71)	kene (59)	mindfulness (67)
8	psalm (77)	pakai (61)	unprecedented (69)
9	revival (79)	alhamdulillah (70)	visualization (75)
10	nak (81)	pasal (74)	defies (79)
11	chc (85)	abeh (76)	corporations (81)
12	fellowship (87)	terawih (82)	hinduism (85)
13	gospel (91)	korang (86)	konjam (89)
14	anointing (93)	cakap (88)	mouthful (93)
15	cornerstone (99)	takpe (92)	une (95)
16	corinthians (105)	ade (96)	bwahahahaha (100)
17	amen (111)	ckp (98)	tomtom (102)
18	hogc (113)	insya (102)	prequel (105)
19	proverbs (117)	tengok (110)	rinpoche (109)
20	churches (119)	takde (112)	composer (111)

Table 4.9: Top 20 SO-CO-NH-BI($\langle word \rangle$)s By SVM Weights (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	pastor (13)	insyaallah (13)	mindfulness (12)
2	church (15)	masjid (15)	meditation (14)
3	hillsong (30)	terawih (30)	dalai (16)
4	chc (32)	tgk (32)	dharma (28)
5	christ (34)	imam (39)	buddhism (30)
6	corinthians (36)	kene (41)	composer (36)
7	prosperous (38)	pulak (48)	visualization (40)
8	christ (40)	kerana (55)	indifferent (44)
9	anointing (42)	madrasah (57)	unprecedented (50)
10	psalm (34)	ade (64)	corporations (52)
11	dismayed (36)	insyallah (66)	defies (59)
12	revival (38)	ustaz (72)	mouthful (61)
13	chc (51)	hijab (80)	une (63)
14	churchill (53)	insya (84)	bwahahahaha (65)
15	hillsong (61)	azan (90)	indifferent (71)
16	teresa (63)	rasulullah (94)	konjam (73)
17	harvest (67)	bukhari (100)	emperor (77)
18	bernard (73)	darul (104)	hinduism (83)
19	forsaking (75)	sentiasa (106)	abrsm (87)
20	pringle (83)	jom (108)	rinpoche (91)

konghee, *JoyceMeyer*, and *JosephPrince* while the followees that Muslims follow are *IslamicThinking*, *IslamSpeaks*, and *Norfasarie*. For Buddhist, the important followees that Buddhists follow are *Zen_Moments*, *Buddhism_Now*, and *elephantjournal* as shown in Table 4.10. These are popular pastors, churches, religion-related users in Twitter.

Table 4.10: Top 20 SO-ST($\langle followee \rangle$)s By SVM Weight (Global Ranks are in Parentheses)

Rank	Christian	Muslim	Buddhist
1	konghee (9)	IslamicThinking (9)	Zen_Moments (5)
2	JoyceMeyer (10)	IslamSpeaks (10)	Buddhism_Now (6)
3	JosephPrince (21)	Norfasarie (21)	elephantjournal (7)
4	MaxLucado (22)	LisaSurihani (22)	DhammaLinks (8)
5	DarleneZschech (23)	MaherZain (23)	theworsthorse (9)
6	JohnBevere (24)	TheNobleQuran (24)	Bodhipaksa (21)
7	CSLewisDaily (25)	syarifsleeq (25)	BuddhistGeeks (22)
8	philpringle (26)	awalashaari (26)	DharmaDots (23)
9	ARBernard (27)	Aaron535Aziz (27)	waylonlewis (24)
10	JoelOsteen (28)	WardinaSafiyyah (28)	tricyclemag (25)
11	hillsongunited (29)	ImranAjmain (29)	zenrev (26)
12	chcsg (44)	MediaCorp_Suria (36)	LamaMarut (27)
13	BrianCHouston (45)	iNasuha (37)	SharonSalzberg (32)
14	nccsg (46)	fadhilfadaroz (38)	dhammagirl (33)
15	joepurcell (47)	didicazli (46)	Blogisattva (34)
16	Celestfoo (48)	MuslimSG (47)	djbuddha (35)
17	thezoneministry (55)	HyrulAnuar (51)	ponlop (46)
18	JianMingTan (56)	DearAbdullah (52)	LamaSuryaDas (47)
19	garrettleejw (65)	Fiza_O_ (53)	shambhalasun (48)
20	Chris_Honegger (66)	TaufikBatisah (54)	thichnhathanh (49)

4.5 Conclusion

From our experiment results, we conclude that collective classification method which uses the top scored unlabeled users as pseudo-labeled users for training is superior than than multiclass classification. By learning from unlabeled users, we actually can improve the classification performance. The experiment further shows that the label-dependent features which change with additional pseudo-labels are discriminative features in region prediction. The feature ranking results further enumerate the important features.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In this thesis, we have proposed two methods to address religion prediction in social networks, namely, the multiclass classification method using content and social structural features, and the collective classification method. The proposed methods make use of both textual features of users and their social structural features. The collective classification method consists of multiple iterations, each iteration adding top scored users of the previous iteration as pseudo-labeled users.

We use a comprehensive set of features to represent multiple aspects of Twitter users. We systematically construct a taxonomy of these features. They are divided into label-independent and label-dependent. Beyond this, we develop interesting approaches that calibrate the importance of users in the religion communities using the religion affiliation of their neighbors.

We evaluate our method on a real data set crawled from Twitter. The data set includes a large set of active Singapore Twitter users. Among them, a small set of users who declare explicitly themselves as the followers of three religions are labeled carefully. An analysis of top followees is presented to show the potential usefulness of using followee features.

The experiments clearly show the benefit of making use of all types of features and the collective classification method. The experiments also show that in the religion classification problem, label-dependent features are more important than label-independent features.

5.2 Future Works

In Chapter 3, we talked about threshold adjustment technique and its benefit. We can further consider distribution of data and user feedback if available to improve threshold adjustment technique. Another direction is to improve SVM learning objective function to consider the data imbalance issues in learning SVM.

Our network follow links are homogeneous meaning there is one kind of them. We can in the future consider to include a categorization of links, dividing them into different categories like family, school/university, work, hobby. We believe such categorization will help user profiling process. However, the dataset of such network links is lacking.

We also can incorporate a richer set of textual features and link analysis features to improve the classification algorithm. The potential of more complex features are promising as they provide deeper view of underlying characteristics.

Bibliography

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

[2] Howard Aldrich and Catherine Zimmer. Entrepreneurship through social networks. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1986.

[3] Chumki Basu, Haym Hirsh, William Cohen, et al. Recommendation as classification: Using social and content-based information in recommendation. In *AAAI/IAAI*, pages 714–720, 1998.

[4] Charles Daniel Batson, Patricia Schoenrade, and W Larry Ventis. *Religion and the individual: A social-psychological perspective.* Oxford University Press, 1993.

[5] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on twitter. In *NAACL*, 2013.

[6] Aditya Bhargava and Grzegorz Kondrak. Language identification of names with svms. In *HLT*, 2010.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[8] Souvik Debnath, Niloy Ganguly, and Pabitra Mitra. Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th international conference on World Wide Web*, pages 1041–1042. ACM, 2008.

[9] Yasser El-Manzalawy and Vasant Honavar. Wlsvm: Integrating libsvm into weka environment. *Software available at http://www. cs. iastate. edu/yasser/wlsvm*, 2005.

[10] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[11] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010.

[12] Brian Gallagher and Tina Eliassi-Rad. Leveraging label-independent features for classification in sparsely labeled networks: An empirical study. In *Advances in Social Network Mining and Analysis*, pages 1–19. Springer, 2010.

[13] Anthony Giddens, Mitchell Duneier, and Richard P Appelbaum. *Introduction to sociology*. WW Norton, 1996.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[15] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62, 2005.

[16] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[17] Thorsten Joachims. Making large scale svm learning practical. 1999.

[18] Sang-Bum Kim, Hae-Chang Rim, Dongsuk Yook, and Heui-Seok Lim. Effective methods for improving naive bayes text classifiers. In *PRICAI 2002: Trends in Artificial Intelligence*, pages 414–423. Springer, 2002.

[19] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, pages 4–15. Springer, 1998.

[20] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[21] Luke K McDowell, Kalyan Moy Gupta, and David W Aha. Cautious collective classification. *The Journal of Machine Learning Research*, 10:2777–2836, 2009.

[22] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[23] Richard J Oentaryo, Ee-Peng Lim, David Lo, Feida Zhu, and Philips K Prasetyo. Collective churn prediction in social network. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 210–214. IEEE Computer Society, 2012.

[24] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *SIGKDD*, 2011.

[25] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. In *ICWSM*, 2011.

[26] Foster Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI2000 workshop on imbalanced data sets*, pages 1–3, 2000.

[27] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Int'l Workshop on Search & Mining User-generated Contents*, pages 37–44. ACM, 2010.

[28] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[29] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[30] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[31] James G Shanahan and Norbert Roma. Improving svm text classification performance through threshold adjustment. In *Machine Learning: ECML 2003*, pages 361–372. Springer, 2003.

[32] Ms Seet Chia Sing, Ms Wong Wei Lin, and Expenditure Income. Singapores census of population 2010. *Statistics Singapore Newsletter*, 6:23–27, 2009.

[33] Aixin Sun, Ee-Peng Lim, and Ying Liu. On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201, 2009.

[34] Aixin Sun, Ee-Peng Lim, Wee-Keong Ng, and Jaideep Srivastava. Blocking reduction strategies in hierarchical text classification. *Knowledge and Data Engineering, IEEE Transactions on*, 16(10):1305–1308, 2004.

[35] Harry C Triandis. The self and social behavior in differing cultural contexts. *Psychological review*, 96(3):506, 1989.

[36] V Vapnik. The nature of 6tatistical learning theory. *Data Mining and Knowledge Discovery*, pages 1–47, 6.

[37] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.