# Offering Language Based Services on Social Media by Identifying User's Preferred Language(s) from Romanized Text

Mitesh M. Khapra
IBM Research India
Bangalore
India
mikhapra@in.ibm.com

Salil Joshi
IBM Research India
Bangalore
India
saljoshi@in.ibm.com

Ananthakrishnan Ramanathan
IBM Research India
Bangalore
India
anandr42@gmail.com

Karthik Visweswariah
IBM Research India
Bangalore
India
v-karthik@in.ibm.com

## ABSTRACT

With the increase of multilingual content and multilingual users on the web, it is prudent to offer personalized services and ads to users based on their language profile (*i.e.*, the list of languages that a user is conversant with). Identifying the language profile of a user is often non-trivial because (i) users often do not specify all the languages known to them while signing up for an online service (ii) users of many languages (especially Indian languages) largely use Latin/Roman script to write content in their native language. This makes it non-trivial for a machine to distinguish the language of one comment from another. This situation presents an opportunity for offering following language based services for romanized content (i) hide romanized comments which belong to a language which is not known to the user (ii) translate romanized comments which belong to a language which is not known to the user (iii) transliterate romanized comments which belong to a language which is known to the user (iv) show language based ads by identifying languages known to a user based on the romanized comments that he wrote/read/liked. We first use a simple bootstrapping based semi-supervised algorithm for identify the language of a romanized comment. We then apply this algorithm to all the comments written/read/liked by a user to build a language profile of the user and propose that this profile can be used to offer the services mentioned above.

## Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

## General Terms

Experimentation


*WWW 2013 Companion,* May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

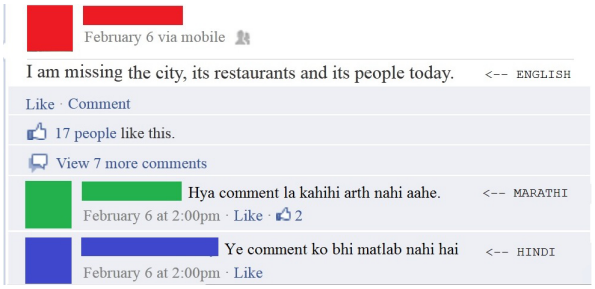


**Figure 1:**

## Keywords

Language profile, Romanized Content, Social Media

## 1. INTRODUCTION

Building user profiles based on their activities on social networking sites and microblogging websites has gained a lot of attention in the recent past[3]. Such profiles help in enhancing user experience by offering personalized services. This, in turn, increases a user's loyalty to a particular service or website. Additionally, such profiling helps service providers to generate more ad revenue by showing better targeted ads to users. In this work, we focus on a particular category of personalized services, *viz.*, language based services depending on the list of languages known to a user and the extent to which he uses them. One may argue that it is trivial to identify the language profile of a user based on his/her geographical location or comments/articles written/read by him. However, there are three factors which make this task non-trivial. First, in multilingual countries (*e.g.*, India) most internet users are bilingual if not multilingual which makes it difficult to identify a user's language profile based on his geographical location. Second, although while registering on social networking sites users are requested to provide a list of languages known to them, it is very hard to ensure that users provide complete/correct

information. Third, many users write regional (non-English) content using Latin/Roman script which makes it hard to distinguish one language from another using script based features.

In this paper, we focus on providing language based services in the presence of the third factor listed above, *viz.* extensive use of Latin/Roman text for writing regional (non-English) content. To motivate a few language based services that can be offered in the presence of romanized content, we point the reader to the example in Figure 1. The figure shows the original post by user $A$ (in English) followed by two comments in two different languages (Hindi and Marathi) but both written using Latin/Roman script. Now, consider a user $B$ who visits the wall of user $A$ and speaks only Hindi and English. User $B$ would be served better if the last comment (in Marathi) is hidden (or translated to English or Hindi). Note that facebook does provide the option of "Translate this post" automatically for some languages if the post is written using the native script of that language but it does not provide this option for regional (non-English) posts written using Latin/Roman script.

## 2. IDENTIFYING LANGUAGE OF ROMANIZED CONTENT

Identifying the language of text written using native script was considered to be a solved problem for a long time[2]. However, recently there has been some renewed interest in this field where [1] have shown that this task becomes challenging if (i) the number of languages is large, (ii) the length of text is small and (iii) there are fewer training instances. Since we are dealing with comments posted on social networking sites and have very limited training data, all of the above conditions are true for our task. In addition, we have the following challenges (i) comments in multiple languages are written using the same Latin/Roman script and (ii) comments on social networking sites are typically very noisy which makes the task even harder. We propose a simple bootstrapping based approach where we start with a small number of labeled posts and train a multi-class SVM (each language is a different class). We use character n-grams and word n-grams as features. This classifier is then used to label a large number of unlabeled posts and the posts which get labeled with a high confidence are added to the training data. This increased data is then used to re-train the classifier and the above process is repeated till convergence.

## 3. IDENTIFYING USER'S LANGUAGE(S)

Once the above classifier is trained we use it to label all the posts that a user wrote or liked. If any of the posts written/liked by a user belong to language $L_i$ then $L_i$ is added to the list of languages known to that user. Further, we use a simple formula to find the *extent* to which a user uses a particular language:

$$usage(L_i) = \frac{number\ of\ posts\ written/liked\ in\ language\ L_i}{total\ number\ of\ posts\ written/liked\ by\ the\ user}$$

## 4. OFFERING LANGUAGE BASED SERVICES

Once the language profile of the user has been built, we can offer him/her language based services. Continuing with the example given earlier, when $B$ visits the wall of $A$ the following language based services can be offered to him/her:

- An option to translate the first post (in Marathi) to one of the 2 languages (Hindi, English) known to him/her.
- An option to re-rank the posts so that the posts written in one of the 2 languages (Hindi, English) known to him/her appear at the top (sorted according to *usage*).
- An option to convert the second post (in Hindi) to Devanagari script for better readability.
- Show Hindi ads to the user.

## 5. EXPERIMENTS AND RESULTS

We collected around 50 posts each in 4 languages, *viz.*, English, Hindi, Marathi and Bengali. We used 80% of this data as the seed data for our bootstrapping algorithm. The remaining 20% was used as held-out validation set. In addition, we collected about 2K unlabeled romanized posts which belonged to several Indian languages. A publicly available (http://svmlight.joachims.org/svm_multiclass.html) multi-class SVM tool was used for training our classifier. The initial classifier trained on the seed data was used to label the 2K posts and the posts labeled with high confidence were added to the training data. We repeated this process for 5 iterations. Using just the seed data we got an accuracy of 84.4% on the validation set, whereas, after running our bootstrapping algorithm for 5 iterations we were able to achieve an accuracy of 91.1%.

Next, we collected posts written/liked by 100 users. We knew the actual language profile of these users since these users were known to us. We then identified the language profile of these users using our algorithm. For 72 out of the 100 users we were able to identify all the languages known to the user and no additional incorrect languages were identified. For 14 out of the 100 users we were able to identify only a subset of the languages known to the user but none of the languages identified were incorrect. Lastly, for the remaining 14 users our algorithm correctly identified all the languages known to the user but also added one or more incorrect languages to the user's language profile.

## 6. CONCLUSION

We proposed some simple algorithms for identifying the language of romanized text and for building a language profile of users on social networking sites. Based on this information we propose to offer language based services to users.

## 7. REFERENCES


[1] T. Baldwin and M. Lui. Language identification: the long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 229–237, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[2] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[3] H. Villiard and M. A. Moreno. Fitness on facebook: Advertisements generated in response to profile content. *Cyberpsychology, Behavior, and Social Networking*, 2012.