Gender Inference of Twitter Users in Non-English Contexts

Morgane Ciot
School of Computer Science
McGill University
Montreal, Quebec, Canada
morgane.ciot@mail.mcgill.ca

Morgan Sonderegger
Department of Linguistics
McGill University
Montreal, Quebec, Canada
morgan.sonderegger@mcgill.ca

Derek Ruths
School of Computer Science
McGill University
Montreal, Quebec, Canada
derek.ruths@mcgill.ca

Abstract

While much work has considered the problem of latent attribute inference for users of social media such as Twitter, little has been done on non-English-based content and users. Here, we conduct the first assessment of latent attribute inference in languages beyond English, focusing on gender inference. We find that the gender inference problem in quite diverse languages can be addressed using existing machinery. Further, accuracy gains can be made by taking language-specific features into account. We identify languages with complex orthography, such as Japanese, as difficult for existing methods, suggesting a valuable direction for future research.

1 Introduction

A 2012 study reported that US-based Twitter users now account for only 28% of all active accounts on the platform (Semiocast, 2012). Brazil, Japan, India, and Indonesia all rank in the top 10, each with over 5% of all users. These and other findings confirm that Twitter enjoys widespread international popularity and usage. This is also reflected in the multinational community of researchers who study human behavior on Twitter and related platforms, e.g. (Sakaki et al., 2010; Tumasjan et al., 2010; Kim and Park, 2012).

It is remarkable, then, that advances in latent attribute inference on social media have been largely confined to English content, e.g. (Liu and Ruths, 2013; Zamal et al., 2012; Pennacchiotti and Popescu, 2011; Conover et al., 2011a). This bias may be partially explained in the context of the research being conducted largely by anglophone researchers. Nonetheless, it has created a notable silence in the literature concerning the large-scale analysis of languages, cultures, and people on social media who do not employ English.

In this paper, we examine the problem of latent attribute inference outside the English-language context. To our knowledge, this is the first such study ever conducted. Here we specifically focus on gender inference, as it has been the basis for significant work in recent years (Liu et al., 2012; Zamal et al., 2012; Pennacchiotti and Popescu, 2011; Rao et al., 2010; Burger et al., 2011). Our work makes two contributions. First, we quantify the extent to which established gender inference methods can be used with non-English Twitter content. Second, we explore the capacity for unique features of other languages (besides English) to improve inference accuracy. This second aspect, in particular, acknowledges the fact that latent attribute inference may be easier in some languages due not to conventions in word usage, but to syntactic structure.

In order to assess the extent to which existing gender inference machinery works for users who use languages other than English, we assembled Twitter datasets for languages that are both prevalent on Twitter and representative of diverse language families: Japanese, Indonesian, Turkish, and French. Each dataset consisted of approximately 1000 users who tweeted primarily in a given language. We used Amazon Mechanical Turk to manually label each user with their gender, using a language-agnostic labeling strategy (Liu and Ruths, 2013). For classification, we employed a performant support vector machine-based (SVM) technique that has been used in a range of studies, e.g. (Rao et al., 2010; Burger et al., 2011; Zamal et al., 2012).

We found that, without any modification to the types of features given to the SVM, the classifier accuracy was comparable on English, French, and Indonesian. Turkish actually performed much better, achieving 87% on average. Gender in Japanese, in contrast, could not be reliably inferred with any reasonable accuracy (61% on average) despite numerous attempts to preprocess the tweets and tune the classifier to accommodate the language's complex orthography. This indicates that existing approaches may not generalize well to language systems with thousands of distinct unigrams (as opposed to tens or hundreds in the other languages considered).[1]

To evaluate the extent to which language-specific features might be used to boost the accuracy of the SVM classifier further, we focused on French. French is a valuable case study because, unlike English, it has a number of syntax-based mechanisms that can encode the gender of the speaker. The most common instantiation of gender marking is the modification of adjective and some past participle endings to match the gender of the subject in constructions beginning with "je suis" (trans. "I am") constructions. A classifier based on this insight achieved average accuracy of 90% on the vast majority of French users, surpassing the accuracy of standard techniques on English or French.

Overall, our results show that, with little modification, existing gender inference machinery can perform comparably to English on several other languages. There are clear areas for substantial improvement: incorporating language-specific features and, in the case of Japanese, finding better ways of accommodating the complex orthography. These findings identify promising directions for future research and will, hopefully, call attention to an important area in the latent attribute inference domain in need of further work.

2 Background

Non-English Twitter data mining and studies. Existing work on non-English Twitter content can been divided into two groups: surveys of the use of several languages on the platform and studies of a social phenomenon in a non-English body of tweets. To our knowledge, only a handful of the former variety exist. One recent paper characterized the relationship between language and geography (Mocanu et al., 2012). Another measured how high-level tweet features (i.e., link, mention, and hashtag frequencies) vary across languages (Weerkamp et al., 2011). These papers show that tweet structure and content can differ widely across languages.

More work has been done in the latter category: analysis of social phenomena in a non-English context. A well-known study evaluated usage of Twitter in the aftermath of the 2010 earthquake in Japan (Sakaki et al., 2010). Another Japanese-oriented study evaluated the impact of television on tweeted content (Akioka et al., 2010). Within this category, another recurring research topic is the analysis of political discussion and elections. Outside of English-based analysis, some attention has been given to European and East Asian elections, e.g. (Tumasjan et al., 2010; Giglietto, 2012; Kim and Park, 2012). However, few of these studies have considered measures beyond simple hashtag frequencies, relative mention counts among politicians, and retweet counts. The only study using more complex features for computational text analysis involved sentiment analysis of a set of German tweets (Tumasjan et al., 2010). However, the tweets in this study (conducted at a German university) were translated into English prior to analysis, a step which underscores the significant bias towards English in the literature on analyzing microtext, and the tools available to researchers in this domain.

Gender inference methods. Gender inference is a field of research situated with the broader area of latent attribute inference. The majority of recent work in this area has focused on Twitter users (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Conover et al., 2011b; Burger et al., 2011; Rao and Yarowsky, 2010; Liu and Ruths, 2013; Liu et al., 2012; Zamal et al., 2012). Classifiers have been built, predominantly, using support vector machines, e.g. (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Burger et al., 2011; Zamal et al., 2012), though boosted decision trees and latent dirichlet allocation systems have also been evaluated, e.g. (Pennacchiotti and Popescu, 2011; Conover et al., 2011b). With one exception, gender inference accuracy has been re-

[1] In this work, *unigram*, *bigram*, and *k-gram* refer to one, two, and k-character sequences in a language's written form.

ported between 80% and 85%. The one study which reported 90% accuracy involved the use of a dataset which has been shown to be quite different from typical anglophone Twitter users (Burger et al., 2011). This same study did involve non-English Twitter users, but did not analyze the performance of the classifier on different languages (e.g. break down performance by language, examine to what extent its results were due to better performance on some languages), or indeed discuss fully which languages were present in their sample. Thus, little can be inferred from Burger et al.'s study about the relative performance of attribute inference methods on different languages, which is the focus of our paper.

Language families. Human languages can be classified into different *language families*, defined as a set of languages which are all descended from a single, ancient parent language. Languages which are genetically related (in the same family), however distantly, tend to share many more characteristics than languages from different families.

Each language considered in this paper belongs to a different language family: French to Indo-European, Turkish to Altaic, Japanese to Japonic, and Indonesian to Austronesian. Thus, these languages are completely genetically unrelated, by definition. Further, they are both geographically and culturally dispersed. While they all have some loanwords from English, these constitute a tiny fraction of each language's vocabulary. This selection of languages allows us to conduct the most far-reaching survey of non-English latent attribute inference performance to date.

A variety of features make each language selected interesting within the gender inference context. French is noteworthy for its grammatical gender. All nouns, including people, are grammatically "male" or "female." English, in contrast, has separate pronouns for people of different genders (e.g., "he", "she"), but does not have *grammatical* gender. (Besides a handful of exceptions like "waiter"/"waitress", there are no words besides pronouns which have different "masculine" and "feminine" forms.) Indonesian, Turkish, and Japanese are all so-called *genderless languages*. Like many languages of the world, they do not have distinct male and female pronouns (like English and French), or grammatical gender (like French).

3 General Gender Inference

In order to evaluate the extent to which existing gender inference machinery can be used on users whose tweets are in languages other than English, we developed gender-labeled datasets of Twitter users for each language and then evaluated the performance of a classifier on each.

3.1 Data

The core data for this project consisted of four datasets of content from Twitter users who tweeted predominantly in one of four languages—French, Indonesian, Turkish, and Japanese—collected using the methods described below.

Data collection. In order to identify users for candidate inclusion in a particular language's (hereafter the *target language*) dataset, we walked the streaming output of the Twitter firehose and evaluated the language of each tweet using the language models provided by the Natural Language Toolkit (Bird, 2006). Users associated with tweets written in the target language were added to a list. 5000 such users were identified for each language. The latest 1000 tweets for each user were downloaded. This comprised the base for the target language dataset.

Assigning gender labels. In prior work, e.g. (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Zamal et al., 2012), the dominant way of obtaining datasets consisting of Twitter users with high-confidence gender-labels is to use gender-name associations. The use of name-gender associations are problematic when non-English content is considered because databases of anglophone name-gender associations are no longer useful (Mislove et al., 2011). We instead used Amazon Mechanical Turk workers to identify the gender of the person shown in the profile picture associated with a user's account (Liu and Ruths, 2013). In our datasets, each user's profile picture was coded by 5 separate workers. Users with non-photographic or celebrity-based profile pictures was discarded, as well as any users with profile pictures where the gender could not be confidently assessed (less than 4 out of 5 votes for one gender).

Table 1 shows the final composition of each dataset. In Japanese and Indonesian, we observed

Table 1: The composition of the different language datasets used in this study.

Language	# Males	# Females	Total Size
French	437	506	943
Indonesian	977	2260	3237
Turkish	1672	1937	3609
Japanese	309	520	829

a notable difference in the number of males and females in the dataset. Measures were taken to ensure that classifier results were not biased by these differences within the datasets.

3.2 Methods

The majority of prior work in gender inference (and latent inference in general) has used support vector machines (SVMs). We followed prior work in this regard, particularly since our intent here is to evaluate the relevance of existing gender inference machinery on other languages. For the present study, we adopted an SVM-based classifier, described in (Zamal et al., 2012), that incorporated nearly all features used in prior work and showed comparable (and sometimes better) accuracy than other methods. Parameter values and kernel choices for the SVM are discussed in the source paper.

Feature set. SVM classifiers require that each object to be classified be represented by a fixed-length feature vector. The features we employed were: k-top words, k-top digrams and trigrams, k-top hashtags, k-top mentions, tweet/retweet/hashtag/link/mention frequencies, and out/in-neighborhood size. Note that "*k-top X* features" (e.g., k-top hashtags) refers to the k most discriminating items of that type for each label (i.e., Male/Female). Thus, *k-top words* is actually $2k$ features: the k words most associated with males and the k words most associated with females.

This list of features is the same set of features used in (Zamal et al., 2012), except that *k-top stems* and *k-top co-stems* were both dropped in our version. Both of these feature types are specific to English. Of course, word stems do exist in other languages, however we found that stemmers (the algorithms that identify and extract the appropriate stem from a word) were not available across the whole bank of languages. Therefore, we omitted these stem and co-stem features. We also added features for the usage frequencies of Eastern-style and Western-style emoticons but saw no discernible change in accuracy; thus, these features are not discussed further.

It is important to note that all features included in our classifier are language-agnostic. An n-gram is simply an n-character sequence drawn from the alphabet and additional symbols (numbers, punctuation, etc.) present in tweets written in the target language. Words are sequences of characters that are bounded by whitespace or punctuation. Hashtags are words proceeded by a pound ("#") character, mentions by an "@" symbol. A system that properly supports unicode strings can implement all of these notions without knowing anything about the target language it is operating on.

Tokenization of Japanese. While all the definitions provided above for the SVM features are *operational*, there is a glaring disconnect between the whitespace-border definition of a word and written conventions in Japanese. Specifically, in Japanese words are generally not separated by whitespace.

We used a tokenizer to insert whitespace into Japanese text to break up words. Tokenization was done using Kuromoji, the software Japanese morphological analyzer used and supported by the Apache software Foundation (Atilika, 2012). Notably, this tool tokenizes the mixed character sets that are often used in informal Japanese writing.

As tokenization does involves some language-specific processing, its use here somewhat undermines the objective set out for this project. Thus, we report the accuracy achieved for both untokenized and tokenized Japanese tweets. Curiously, tokenization was found to not make a difference in overall average accuracy.

3.3 Results

For each dataset, 5-fold cross validation was used to assess the classifier's performance. The value of $k = 20$ was used for all k-top features, though the results reported are robust to changes of this value within reason (between 10 and 30). If the numbers of male and female users were unbalanced in a dataset, the larger set was subsampled randomly to obtain a set of users the same size as the smaller labeled set. During the training process, the actual

Table 2: The accuracy of the SVM-based classifier on each of the language datasets. In the case of Japanese, the performance is given for both the tokenized and untokenized versions of the dataset. (Note that tokenization did not affect overall accuracy.)

Language	Male	Female	Overall
French	0.79	0.73	0.76
Indonesian	0.87	0.80	0.83
Turkish	0.89	0.85	0.87
Japanese (t)	0.50	0.76	0.63
Japanese (u)	0.58	0.68	0.63

values of the features were extracted from the training users (e.g., the k-top differentiating words for males and females were identified). In this way, the gender model implemented by the SVM *was* language-specific, in the sense that a particular language's gender model contained a different set of features. We call our method language-agnostic on the grounds that, given a labeled set of users and tweets drawn from a particular language, a model can be built without any knowledge of the structure or content of the language itself.

Tables in Supplementary Material show the features for the classifier built over each language's entire dataset. Note that to conduct the cross-fold evaluation, new models (and hence different features) were recomputed for each fold. As a result, the features reported are slightly different from those that might have appeared in the models for a given fold. Manual inspection, however, revealed that differences were slight. The features reported in the Supplementary Material can be safely considered a consensus among the models for the individual folds.

The accuracy of the classifier for each language is shown in Table 2. Overall, the classifier demonstrated good performance on all languages except for Japanese. Below, we consider the results for each of the four languages in turn. In each case we discuss language-specific trends in which words were most informative for inferring user gender, and thus help explain the classifier's performance. Throughout, we omit discussion of non-alphanumeric "words" (such as punctuation or emoticons), and call the k-top discriminating words for male and female users the *k-top male words* and *k-top female words*.

French. The k-top words for men and women are of very different grammatical types. Most male words are prepositions or articles (16/25; e.g. *de* 'of', *un* 'a/one'); a few others are basic grammatical words (*ne* '[part of] not', *et* 'and'), or pronouns or verb forms referring to a single person or object (he/she/it), as well as one noun (*France*). In contrast, many female words (11/25) are pronouns or basic verb forms referring to the speaker or a single addressee (*je* 'I', *mon*/*mes*/*ma* 'my', *tu* 'you', *j'ai* 'I have'). Others are pronouns or basic verbs refer to a single person or object (*elle*, 'she/it', *c'est* 'it's'), as well as a few other frequent words (*trop* 'too much', *pas* '[part of] not', *oui* 'yes'). The most salient pattern is that use of words (pronouns, basic verbs) associated with talking about the speaker or addressee indicates a tweet is more likely to be from a female user. Heavy use of other common function words, specifically prepositions and articles, suggests a male user. These patterns reflect known gender differences in word usage by male and female French speakers (Witsen, 1981).

Indonesian. Indonesian achieved performance closest to the inference accuracy for English reported in the literature. The k-top lists for men and women give some justification for why the classifier performed well. Some differences can be tentatively linked to general trends in how men and women use language differently across cultures. 5/25 of men's k-top words are nouns which are either related to soccer (*vs* 'versus', *chelsea* '[name of UK soccer team]', *pemain* 'player') or which could be related to soccer (*jakarta*, *indonesia*, *malam* 'night'); in contrast, no women's words are nouns. It seems plausible that men tweet about soccer significantly more than women. In such a situation, a reasonable concern is that our classifier discriminated soccer from non-soccer enthusiasts rather than males from females. To address this, we confirmed that these topic-based words were not required for accurate classification: a classifier in which soccer words were explicitly removed performed just as well (83.8% vs. 83.3%).

More interestingly, many of the k-top words correspond to men and women using different terms of address and self-reference. Among the k-top words, 7/25 for men and 4/25 for women are terms

of address or self-reference. The terms men use are mostly highly informal, including the slang term *lu* (you) and the English borrowing *bro*; the address terms women use are mostly medium-formality, such as *aku* (I) and *kamu* (you). Thus, women seem to be using "more polite" self-reference and address terms than men on average on Indonesian Twitter, in line with the more general tendency for women to use polite forms more frequently than men cross-culturally (Holmes, 1995).

Turkish. Turkish achieved notably high accuracy: the highest of all four languages considered. In fact, to our knowledge, this is the highest accuracy achieved in the entire Twitter gender inference literature on a dataset drawn from the Twitter general population. The k-top lists of male and female words again give some justification for the classifier's performance. Many differences between the male and female lists can be linked to men and women talking about different topics, or to different people. Several of the male words refer to soccer (*gol* 'goal', *galatasaray* 'popular Istanbul team', *maç* 'match', *at* '[part of imperative for] score'), which men plausibly tweet about more. As with Indonesian, a concern is that topics represent a biased sample of the population. Thus, we tested a classifier with soccer-specific terms removed, and again found no difference in accuracy (86% vs. 87%). Many other k-top words are familiar terms of address for men (*lan*, *abi*, *karde sim*, *adam*, *kanka*) or a greeting used mainly between men (*eyvallah*), suggesting that male users are addressing or discussing men more often than female users are. In contrast, 9/25 of the k-top female words are pronouns referring to the speaker, a familiar addressee, or a third party (he/she/it), while none of the k-top male words are, suggesting female users are more often talking directly about themselves or to others. Finally, 2/25 of the k-top male words are profanity (*amk*, *ulan*), while none of the female k-top words are, suggesting male users swear more.

Japanese. Beyond the Japanese classifier's generally poor accuracy, it is striking that tokenization did not improve overall accuracy. This indicates that once words were properly tokenized, no additional gender-distinguishing signal could be extracted. This may be an indication that word-based features carry little information in languages with complex orthography, such as Japanese (with many thousands of unigrams).

Despite the classifier's poor performance, the k-top discriminating words for male and female users differ in interesting ways. Some differences can be understood as resulting from known general trends in how Japanese men and women's use language. Japanese speakers have a choice of many first-person singular pronouns (equivalent to "I"), which signal different levels of politeness and of male versus female speech. The pronoun *boku* (僕) is associated with informal male speech; accordingly, it is among the k-top male words. Japanese also uses an extensive system of verb forms corresponding to different levels of politeness, and honorifics (affixes for names used when referring to others). Women tend to use polite verb forms and honorifics more frequently than men in Japanese speech (Peng, 1981). In agreement with this pattern, several polite verb forms (*-masu*, *-mashi*) and a polite honorific (*o-*) are among the k-top female words, as is a diminutive honorific often used to refer to women (*-chan*).

4 Language-specific Features and Inference

While the classifier performed well across a diverse set of languages, recall that all features used by the SVM were language-agnostic. A natural question concerns the extent to which language-specific features relevant to the attribute of interest (e.g., gender) might improve the classifier's performance.

We examine this question within the context of French. Where gender inference is concerned, French is quite interesting because information about the gender of nouns (including the speaker) is often obligatorily marked in the syntax: many words have different 'masculine' and 'feminine' forms for referring to male and female nouns, including the speaker. Thus, it is in principle often possible to infer the gender of the speaker by which form they use, although it is not clear a priori that this method will work for Twitter data.

4.1 Method

French grammar dictates that which forms of words are used often reflects the gender of the speaker.

Adjectives and past participles all have masculine and feminine forms, which are often spelled differently, and in addition often pronounced differently.Adjectives must agree in gender with the noun they refer to. For example, "I am happy" would be *je suis heureuse* for a female speaker and *je suis heureux* for a male speaker (literally "I-am-happy"); *heureuse* and *heureux* are the feminine and masculine singular forms of the adjective, and are pronounced differently. Past participles of verbs also agree with the gender of the subject or object of the verb, for certain verbs and constructions. For example, "I went" would be *je suis allée* for a female speaker and *je suis allé* for a male speaker (here *suis* is used to form the simple past of the verb *aller*, 'to go'); *allé* and *allée* are the masculine and feminine forms of the past participle of *aller*, and are pronounced the same.

Note that the phrase *je suis* ("I am") occurs in both the adjectival and verbal constructions referring to the speaker; however, the function of *suis* differs between the two. *suis* is the first-person singular form of the verb *être* ("to be"), and functions as a copula when followed by an adjective ("I am happy") but as an auxiliary verb to mark the past tense, when followed by the past participle of certain verbs ("I went"). For our purposes, what is important is that, in both cases, a following adjective or past participle will take on the gender of the speaker.

When this construction occurs in a tweet, it is likely that *je* is referring to the author of the tweet, and the rules of French grammar dictate that the gender of the associated adjective or past participle should reflect the gender of the tweet's author. We implemented a classifier that used this logic to classify the gender of francophone Twitter users. It is worth emphasizing that the *existence* of adjectives and participles which reflect the speaker's gender does not automatically make gender identification in French tweets a trivial task. First, given the prevalence of non-standard spelling and grammar on Twitter and other online platforms, French users may sometimes not use the 'correct' gender marked form reflecting their actual gender—especially given that the male and female forms for a given adjective or participle are often pronounced the same. Second, even if gender-marked constructions are used correctly, they may not occur sufficiently often in tweets to be a reliably used for speaker gender identification. Both of these concerns are borne out in our French dataset, as described further below; the question addressed in the experiment is how useful the signal provided by gender-marked forms is, despite these two sources of noise.

Table 3: The set of patterns that were considered to be suis-constructions when encountered in a tweet.

jn suis pas, jm suis, jmsuis, jnmsuis pas, jnsuis pas, je ne suis pas, je suis pas, jsuis, jensuis pas, jemsuis, jnesuis pas, jmesuis, je me suis, je ne me suis pas

Unlike the probabilistic SVM classifier, the suis-construction classifier can be made entirely deterministic. For a given user, the set of tweets containing a suis-construction are identified, $T_{suis}(u)$. Of these, we can identify the number of those tweets that involve an adjective or past participle with a female ending $T^F_{suis}(u) \subseteq T_{suis}(u)$. Labeling a user involves selecting a threshold based on $T^F_{suis}(u)$ and $T_{suis}(u)$ below which a user receives one label and above which the user receives the other label.

Detecting suis-constructions. As expected, cursory inspection of tweets revealed that Twitter users often employed shorthand forms of the suis-construction. We accounted for this by conducting a manual survey of the shorthand forms of the suis-construction. A catalog of regular expressions was drawn up that matched the different suis-construction forms we identified, shown in Table 3.

Recognizing the gender of the adjective or past participle involved in a suis-construction required a second processing stage. The Lexique lexical database was used to tag the word trailing the suis-construction (New and Landing, 2012). If the tag was not an adjective or verb, the construction was discarded as it would not contain a gender indication. If the word was recognized as an adjective or verb, Lexique would also return the gender, which would be returned as the gender indication for that particular suis-construction.

Threshold selection. We evaluated a number of policies for assigning the user's gender based on the relative values of $T^F_{suis}(u)$ and $T_{suis}(u)$. In the end, however, the best performing threshold was $T^F_{suis}(u) > 1$: simply labeling as female any user

Table 4: The component-wise and overall accuracy of the combined suis-construction and SVM classifier.

Component	# users	Male Acc	Female Acc	Overall Acc
suis-const.	723	0.91	0.90	0.90
SVM	220	0.70	0.54	0.62
Overall	943	0.86	0.82	0.83

who employed the female construction even once. This threshold makes sense given the plausible intuition that females will (almost always) be the only users to employ a female suis-construction; however, it is quite sensitive to uses of female suis-constructions by males.

Mixed classifier. Since not all users had tweets which contained suis-constructions, we combined the SVM-based classifier used previously with the suis-construction-based classifier. The SVM component was applied to any users who lacked suis-constructions entirely in their tweet history. Any user who used even one suis-construction would be labeled according to the $T^F_{suis}(u) > 1$ threshold.

4.2 Results

We ran our classifier on the French dataset, obtaining the results shown in Table 4.

Coverage of the suis-construction. In spite of our concerns over the occurrence frequency and detectability of the suis-construction in tweets, our results show that suis-constructions were found in tweets belonging to nearly 75% of all users in the dataset. This suggests that the suis-construction classifier has quite broad coverage of the population. Of course, given the essential role of the verb "être" in French (like the role of "to be" in English), its frequent use is expected. Nonetheless, the flexible use of grammar and spelling in Twitter and other online contexts raised a genuine concern that occurrences of the suis-construction might not be detected. In fact, when we looked through the tweets of users who were flagged as not having useful suis-constructions in their tweets, we discovered that many actually did. The issue was that they employed highly irregular spellings that our implementation was not able to pick up. Thus, with additional refinement, it may be possible to improve the suis-construction coverage further, well beyond 80%.

Performance of the suis-construction classifier. On the set of users for which the suis-construction was detected, the classifier did very well, achieving an average accuracy of 90%. Recall that the threshold used to generate the results in Table 4 was $T^F_{suis}(u) > 1$. We tested other (larger) thresholds and found that the performance of the method dramatically and monotonically decreased. This was largely due to female users being misclassified as males, indicating that females do not exclusively use female suis-constructions (this was confirmed via manual inspection of a number of female tweet histories). This is different from males, most of whom are quite strict about using only male suis-constructions. Since forming the female form of an adjective or participle typically requires adding an additional character (or more) to the base of the word, this may reflect a tendency towards dropping gender modifiers in favor of typing less.

Performance of the SVM classifier. While the suis-construction classifier performed well, the SVM component did not do nearly as well on the Twitter users that could not be labeled using the suis-construction, achieving an average performance of 62%. At this level of accuracy, the classifier is performing barely better than a random classifier, which would have achieved around 50% accuracy on the label-balanced testing data. This result stands in opposition to our earlier finding that French users could be labeled with 75% accuracy. This disparity suggests that the non-suis-construction users comprise a particularly difficult-to-classify group.

The suis-construction as a filter. The finding that the SVM classifier performed poorly in the combined classification setting suggests that the suis-construction classifier is acting as a very effective filter for users that are hard for it to classify. While we might have preferred better classification accuracy all around, this result is still interesting and useful. Such filters can decrease classification error by simply flagging those users who cannot be easily classified, leaving them to be handled more carefully by more powerful classifiers or human coding. This is precisely the function that the suis-construction classifier appears to play (in addition to classifying the

other users).

This result suggests a question for future work: whether it is possible to build classifiers that accurately label the sets of users that are discarded by the suis-construction classifier.

Performance of the combined classifier. Despite the relatively poor performance of the SVM component, the accuracy of the combined classifier improved on the original SVM-only classifier by 8%, which is a substantial increase in accuracy. With some additional focus on classifying the difficult users who could not be labeled by suis-construction usage, we feel that this accuracy can be increased upwards of 90%.

5 Discussion

In this project, we have extended, for the first time, the latent attribute inference problem to users who tweet primarily in languages other than English. Our study offers several notable insights.

Existing approaches generalize. While accuracy levels certainly vary across languages, overall an existing SVM-based classifier, when trained on users from a given language, can classify the gender of other users from that same language with accuracy comparable to performance reported for English. We suspect that this result will generalize to the inference of other demographic characteristics (e.g., age and political orientation), though this must be explored in future work.

Complex orthography creates unique issues. Japanese stands out as being utterly unclassifiable using existing SVM-based approaches and feature sets. Even efforts to bridge some of the orthographic disconnects between the Japanese language and the assumptions made by the SVM failed to improve performance. This stands out as a clear direction for future work, particularly since apparent issues with the large number of unigrams used by Japanese will create issues for handling (Mandarin) Chinese, the world's most-spoken language.

Language-specific features boost performance. While unsurprising that customizing a classifier to the peculiarities of a given language boosts performance, our use of the suis-construction in French highlights how particular linguistic features may be uniquely well suited to the inference of particular attributes. The results obtained for French stand in contrast to various, relatively unsuccessful attempts to boost gender inference by incorporating syntactic features of English into the classifier (e.g., using stems and co-stems). It seems that some languages have features better suited for certain classification tasks. Identifying and leveraging such features will be an interesting and fruitful direction for future work.

Classifiers as a linguist's tool. In each language, a number of the k-top words align with or suggest gender-specific conventions in that particular language. That a language-agnostic classifier provided such insights highlights its potential for exploring language-specific word usage patterns and nuances. For example, sociolinguistics (a subfield of linguistics) has long studied the different ways men and women use language, especially in spontaneous speech (Eckert and McConnell-Ginet, 2003); recent work has begun to examine how language is used differently by men and women online as well (Bamman et al., 2012). Such studies could be radically scaled up in terms of the number of languages considered using a language-agnostic gender classifier.

6 Conclusion

Though there has been relatively little investigation into latent attribute inference outside of English-language content, we consider it both a fruitful and important area for future research. Here, we have evaluated the capacity for existing inference methods to be used outside their intended English-language context. Furthermore, we have shown how language-specific features might be incorporated in order to boost classifier accuracy further. The positive results suggest that latent attribute inference in the non-English context as a research direction worthy of further attention.

7 Acknowledgements

The authors gratefully acknowledge three anonymous reviewers whose feedback improved the clarity and correctness of the manuscript. The study was supported by grants from the Social Sciences

and Humanities and Natural Sciences and Engineering Research Councils of Canada (SSHRC Insight Grant #435-2012-1802 and NSERC Discovery Grant #125517855) and the Public Safety Canada Kanishka Program.

References

S. Akioka, N. Kato, Y. Muraoka, and H. Yamana. 2010. Cross-media impact on Twitter in Japan. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*.

Atilika. 2012. Kuromoji morphological analyzer. http://www.atilika.org.

D. Bamman, J. Eisenstein, and T. Schnoebelen. 2012. Gender in Twitter: Styles, stances, and social networks. arXiv preprint arXiv:1210.4567.

S Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*.

J.D. Burger, J. Henderson, and G. Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011a. Predicting the politial alignment of Twitter users. In *Proceedings of the International Conference on Social Computing*.

M.D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F Menczer, and A Flammini. 2011b. Political polarization on Twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

P. Eckert and S. McConnell-Ginet. 2003. *Language and gender*. Cambridge University Press, Cambridge.

F. Giglietto. 2012. If likes were votes: An empirical study of the 2011 Italian administrative elections. In *Proceedings of the International Conference on Weblogs and Social Media*.

J. Holmes. 1995. *Women, men and politeness*. Longman, London.

M. Kim and H.W. Park. 2012. e-measuring Twitter-based political participation and deliberation in the South Korean context by using social network and Triple Helix indicators. *Scientometrics*, 90(1):121–140.

W. Liu and D. Ruths. 2013. What's in a name? Using first names as features for gender inference in Twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium*.

W. Liu, F.A. Zamal, and D. Ruths. 2012. Using social media to infer gender composition from commuter populations. In *Proceedings of the When the City Meets the Citizen Worksop*.

A. Mislove, S. Lehmann, Y.Y. Ahn, J.P. Onnela, and J.N. Rosenquist. 2011. Understanding the demographics of Twitter users. In *Proceedings of the International Conference on Weblogs and Social Media*.

D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra, and A. Vespignani. 2012. The Twitter of Babel: Mapping world languages through microblogging platforms. *ArXiv e-prints*, December.

B. New and C. Landing. 2012. Lexique 3. http://www.lexique.org/telLexique.php.

F.C.C. Peng, editor. 1981. *Male/female differences in Japanese*. The East-West Sign Language Association, Tokyo.

M. Pennacchiotti and A.M. Popescu. 2011. A machine learning approach to Twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

D. Rao and D. Yarowsky. 2010. Detecting latent user properties in social media. In *Proceedings of the NIPS workshop on Machine Learning for Social Networks*.

D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the International World Wide Web Conference*.

Semiocast. 2012. Brazil becomes the 2nd country on Twitter, Japan 3rd, Netherlands most active country. http://semiocast.com/publications/2012_01_31_Brazil_becomes_2nd_country_on_Twitter_superseds_Japan.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International Conference on Weblogs and Social Media*.

W. Weerkamp, S. Carter, and M. Tsagkias. 2011. How people use Twitter in different languages. In *Proceedings of the Web Science Conference*.

R. Schenk-Van Witsen. 1981. Les différences sexuelles dans le français parlé: une étude-pilote des différences lexicales entre hommes et femmes. *Langage et société*, 17(1):59–78.

F.A. Zamal, W. Liu, and D. Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International Conference on Weblogs and Social Media*.