Gender Prediction of Indian Names

Anshuman Tripathi
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India 721302
Email: anshu.g546@gmail.com

Manaal Faruqui
Department of Computer Science and Engineering
Indian Institute of Technology
Kharagpur, India 721302
Email: manaal.iitkgp@gmail.com

***Abstract*—We present a Support Vector Machine (SVM) based classification approach for gender prediction of Indian names. We first identify various features based upon morphological analysis that can be useful for such classification and evaluate them. We then state a novel approach of using n-gram-suffixes along with these features which gives us significant advantage over the baseline approach. We believe that we are the first to use n-grams of suffixes instead of the whole word for predictor systems. Our system reports a top F_1 score of 94.9% which is expected to improve further with increase in training data size.**

I. Introduction

Gender Identification of names is an important pre-processing step for many tasks in Artificial Intelligence (AI) and Natural Language Processing (NLP). It can lead to improvement in performance of applications like Co-reference Resolution, Machine Translation, Textual Entailment, Question Answering, Contextual Advertising and Information Extraction. As is often the case for NLP tasks, most of the work has been done for English names. The presently available softwares for gender identification of names work on dictionary look-up methods. To our knowledge, at this time there is no freely available gender identification system available for research purposes.

SVM based classification approach finds use in large number of Machine Learning applications and is generally easier in implementation and better in performance than other classification approaches. We use the SVM library, LIBSVM [1] provided in MATLAB for carrying out our experiments.

Our main contributions lie in the extensive analysis of various word-level features of Indian names which distinguish between the two genders, identifying the features which are most helpful in classification and presenting a state-of-the-art method for gender identification using a Support Vector Machine (SVM) based classification approach.

II. Related Work

SVM based classification has previously been used for language identification of names [2] and has performed better than language models. Reference [2] has used the *n-grams* of words and word length as features and has shown that the classification accuracy increases with n. However, they do not use any other morphological information of words. Gender identification of Chinese e-mail documents [3] used format features, linguistic features and structural features of e-mails in SVM for classification. It concentrates more upon the overall document structure and less on the individual named-entity. SVM has been used for Gender identification from many other media such as images [4], gait recognition [5] and speech signals [6]. To the best of our knowledge, much work has not been done in using SVM classifiers for gender identification of names represented in text and thus there is a need to explore the applicability and analyze the performance of SVM classifiers on textual data.

III. Support Vector Machines

A Support Vector Machine performs classification by constructing an N-dimensional hyper-plane that optimally separates the data into two categories. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data-points of any class, since, in general, the larger the margin, the lower the generalization error of the classifier. Kernel functions are related to the transformation function, used to obtain the feature vector y in the transformed feature set from the feature vector x in the original feature space. Kernel functions are preferred for these transformations to make the final classifier computationally efficient. A transformation function $\varphi(x)$ is related to the corresponding kernel function $K(x, y)$ (if it exists) by the relation:

$$\varphi(x).\varphi(y) = K(x, y) = f(x, y) \tag{1}$$

Where x and y are the feature vectors in the original feature space. Note that the new feature space is of higher dimension (say d') than the original feature space (say d); kernel function thus facilitates the computation of the dot product $\varphi(x).\varphi(y)$ in higher dimension by computing $f(x.y)$ from the dot product $x.y$ in the original space of lower dimension thereby improving the efficiency. Kernel functions also facilitate easy implementation of soft margin and hard margin classifiers.

The two commonly used kernel functions by the SVM classifiers are polynomial and radial basis kernel functions. Since kernel functions are related to the transformation functions, they decide the dimension of the transformed feature space. Increasing the dimension of the new feature space may result in over-fitting on small training data-set. To train an SVM with a kernel function the number of training examples required (so that the classifier is probably approximately correct) increases exponentially with the dimension of the new feature space

 978-1-4244-8943-5/11/$26.00

(decided by the degree of the kernel function used). This effect is called the curse of dimensionality [7].

IV. Dataset

In most of the countries, a person's name is not a characteristic of his place of birth. However in India, the names of people coming from a particular part of the country show similarity. Different lists are available of North-Indian, South-Indian and East-Indian baby names on the internet. We took an almost equal proportion of these names and formed a list containing around 2000 names which were tagged "male" and "female".

The initially compiled data sets contained names having more than one probable spelling. In such cases, to make our system robust, we took all the possible spellings of the word. For example, "Abhijit" & "Abhijeet" both were put up in the training data. A preliminary overview of the composed data showed that all names had length ≥ 4 and contained an almost equal number of Gujarati, Punjabi, Bangla, Hindi, Urdu, Tamil and Telugu names.

Our compiled training data contained 890 female and 1110 male names. Then we compiled our test data from a different website in such a manner that there was no common name in the training and test data. The test data contained 217 names of which 89 were female and 128 male.

V. Morphological Analysis

Names of males and females exhibit very subtle differences. These features are mostly due to the morphological and phonological structure of the name. The linguistic and phonological analysis of North American names [8] enlists a number of such features, a subset of those has been chosen by us for understanding the typical characteristics which distinguish between male & female Indian names:-

- *Vowel ending*: Names of females generally end in a vowel while that of males in consonants. a, e, i, o, u comprise the set of vowels.
- *Number of syllables*: A syllable is a unit of pronunciation uttered without interruption, loosely a single sound. Female names tend to have more number of syllables than males.
- *Sonorant consonant ending*: A sonorant is a sound that is produced without turbulent airflow in the vocal tract. Hindi possesses eight sonorant consonants [9]. Compared to females, male names generally end with a sonorant consonant.
- *Length of the word*: Even though length of a name does not relate to its gender but our analysis showed that males generally have longer names than females.

Table I shows the distribution of the occurrence of these features across our training data. The syllable identification in words was done manually by students who are native speakers of Hindi.

A striking difference between Indian and American names is shown here by the sonorant ending feature. While [8] reports that the percentage of sonorant ending male names is 19% and for females it is 28.3%, our analysis shows that among Indian names, 32.4% of males and only 3% of the female names show the above feature. Also, 96.6% of Indian female names have vowel ending as compared to 60.4% of American female names. The average number of syllables per word for Indian names is almost twice that of the American names. These differences in the word-structure of Indian and American names indicate a need of separate analysis of Indian names.

Henceforth, vowel ending, average number of syllables, sonorant consonant ending and average length of word would be represented by *isVowel*, *numSyll*, *isSonorant* and *lenWord*.

TABLE I
Structure of Indian Names

Features	*Male*	*Female*
isVowel	96.6%	22.81%
numSyll	2.94%	2.64%
isSonorant	3%	32.4%
lenWord	7.00	7.56

TABLE II
Performance of individual feature-trained classifier

Features	F_1 *Score (%)*
isVowel	91.7
numSyll	62.2
isSonorant	59.9
lenWord	55.3
1-gram	71.9
2-gram	80.6
3-gram	71.4

VI. Experiments

A. Possible Features

As stated in the previous section female names differ from that of males in terms of the *numSyll*, *lenWord*, *isVowel* and *isSonorant*. On one hand, we have features like *numSyll* & *lenWord* which do not differ a lot for the two categories and on the other hand, the percentage of words showing *isVowel* and *isSonorant* features vary largely across the two categories. This gives us an idea that *isVowel* and *isSonorant* are the two features which may primarily help in classifying a name.

As suggested by [2] we include n-gram features as well for our analysis. Including n-gram features would try to identify the set of alphabets which occur together frequently as prefixes, postfixes or in between the word in male and female names. Since all names in our training data set had length ≥ 4 we chose *1-gram*, *2-gram* & *3-gram* features in our experiments. We do not include *4-gram* feature as it may lead to over-fitting on the training data and processing it is computationally much more expensive than *n-grams* of lower degree.

TABLE III
TRAINING ON MULTIPLE FEATURES

isSonorant	numSyll	lenWord	isVowel	F_1 *Score (%)*
1*	1	0†	0	73.2
1	1	1	0	79.7
1	1	1	0	89.8

TABLE IV
PERFORMANCE OF (*N-gram*, *isVowel*) TRAINED CLASSIFIERS

Training Size (No. of Names)	F_1 score (%)		
	1-gram	*2-gram*	*3-gram*
500	85.1	85.2	83.4
800	88.5	86.3	82.5
1000	89.4	88.0	83.4
1500	89.8	88.5	84.2
2000	89.8	89.0	84.0

B. Evaluation of features

First, we simply train our system on different sizes of training data varying from 500 to 2000 examples using only one feature at a time and record the best performance shown by every individual feature. According to the results shown in Table II, while *isVowel* comes out to be the strongest; *isSonorant* & *lenWord* appear to be the weakest predictors of gender.

The performance of *isSonorant*, *lenWord* and *numSyll* is close to 50% which is markedly poor, since for classification involving only two classes, a system which assigns a fixed class to each entity would also have a score $\approx 50\%$. Thus we train our system together on these three features and observe an increase in performance as shown in Table III, but none of the combinations could surpass the score achieved by *isVowel*. Other feature combinations performed worse than the results shown in Table III and hence we have not included those results in this paper. The combination of *n-gram* features with *isSonorant*, *numSyll*, *lenWord* and *isVowel* did not perform better than the former four taken together.

Next, we trained our system on n-grams and *isVowel* for different size of training data and observed that the combination of (*1-gram*, *isVowel*) and (*2-gram*, *isVowel*) show an almost linearly increasing performance whereas the performance of (*3-gram*, *isVowel*) is oscillatory and is not linear with increase in size of the training data. Table IV lists the F_1 score obtained with these features using a linear kernel.

C. *N-gram-suffix* feature

The high improvement observed in all the above experiments due to the introduction of *isVowel* feature indicates that a lot more information about the gender of the Indian names can be extracted from its suffix. This motivated us to look solely at the *n-gram* of the suffix of each word instead of taking all the *n-grams*. For example, *1-gram-suffix* means the last letter of the word is a feature, *2-gram-suffix* means the last 2 letters of the word is a feature and so on and so forth.

The dimension of feature space is greatly reduced by only considering the *n-gram-suffix* features, for instance for *3-grams* the dimension reduced from 26^3 = 17,576 to just 395, since only 395 unique *3-gram-suffix* were present in the training data. This reduction in dimension of feature space allows us to consider even *4-gram-suffix* as a feature. For names in the test data which possess an *n-gram-suffix* which is not present in the training data, all the elements in the *n-gram-suffix* feature vector would be zero and its gender would be determined solely by its *isVowel* feature. Thus, the gender of a name, whose *n-gram-suffix* is unknown to the training data, can be determined with 91.7% probability as evident from Table II.

TABLE V
PERFORMANCE OF (*N-gram-suffix*, *isVowel*) TRAINED CLASSIFIERS

Training Size (No. of Names)	F_1 score (%)			
	n = 1	**n = 2**	**n = 3**	**n = 4**
500	92.6	93.1	93.1	94.5
800	92.6	93.1	94.0	94.5
1000	92.6	93.1	93.5	94.5
1500	92.6	94.0	**94.5**	94.0
2000	92.6	94.0	94.5	94.5

As expected, all the results shown in Table V are better than the result obtained by using only *isVowel* as a feature. Hence, *n-gram-suffix* & *isVowel* features together lead to an improvement in the performance of the system. The performance of the classifier trained using *1-gram-suffix* do not change with increase in the amount of training data because the small dimension of feature space leads to an early saturation of the learning algorithm and no new pattern can be learnt from more data. Although, the performance of classifiers trained on *2-gram-suffix* and *3-gram-suffix* show an increase in performance with the increase in training data, the classifier trained on *4-gram-suffix* performs worse as the training data size is increased, this is attributed to the over-fitting of classifier on the training data. The most ideal improvement in learning is shown by *3-gram-suffix* whose performance increases with the increase in amount of data and gets the highest F_1 score.

D. RBF kernel

The use of Radial Basis Function (RBF) as Kernel function has been found to work well for a wide variety of applications. An RBF is a real valued function whose value depends on the distance from some other point x_j.

$$\varphi(x_i) = e^{-\gamma(x_i - x_j)^2}, \ where\ \gamma > 0 \qquad (2)$$

Since RBF has infinite dimensions it is expected to fit better on the training data. Experiments carried out using RBF as kernel show inferior performance as compared to the linear kernel. Figure 1 shows the performance of the classifier using RBF and Linear functions as kernel on *3-gram-suffix* & *isVowel* features. The worse performance is likely to be caused because of over-fitting of the classifier on the training data.

* '1' means presence of feature † '0' means absence of feature

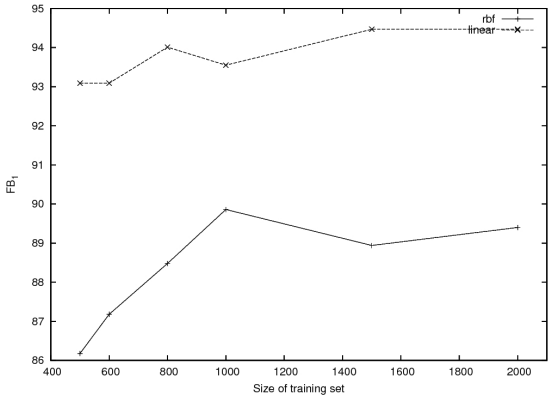

Fig. 1. Performance of RBF kernel function with *3-gram-suffix* feature

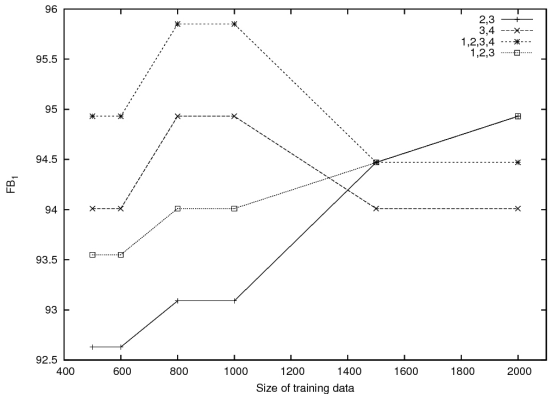

Fig. 2. Performance of different (*n-gram-suffix*, *isVowel*) combinations

E. Combination of n-gram-suffix features

Reference [2] uses a combination of *n-grams* from n = 1 up to some specified length and reports an increase in performance of language identification as the value of n is increased. We exhaustively experimented with different combinations of *n-gram-suffix* features trained on different sizes of training data and present the best results obtained in Table VI. The earlier argument that *3-gram-suffix* is the most ideal feature for gender prediction is further strengthened by its presence in all the best performing *n-gram-suffix* combinations. From Figure 2, it can be seen that while the feature combination of n = 1, 2, 3, 4 achieves the highest score of 95.8% on test data, n = 1, 2, 3 shows a gradual increase in performance with increase in size of the training data and reaches a maximum of 94.9%. This score is expected to increase further with a larger training set. All feature combinations which include *4-gram-suffix* achieve a local maxima value and then decrease and become constant. As stated earlier this phenomenon of decrease in performance occurs due to the probable over-fitting of the classifier on the training data. Thus, we conclude that the feature combination of n = 1, 2, 3 along with *isVowel* is the best predictor for gender of Indian names.

TABLE VI
PERFORMANCE OF (COMBINED *N-gram-suffix*, *isVowel*) TRAINED CLASSIFIERS

Training Size (No. of Names)	F_1 **score (%)**			
	n = {3,4}	**n = {1, 2, 3}**	**n = {2, 3, 4}**	**n = {1, 2, 3, 4}**
500	94.0	93.5	94.0	94.9
800	94.9	94.0	94.9	95.8
1000	94.9	94.0	94.9	95.8
1500	94.0	94.5	94.0	94.5
2000	94.0	**94.9**	94.0	94.5

VII. CONCLUSION

We have presented a study on gender prediction of Indian names using a Support Vector Machine based classification approach. Our study has shown the differences between the structure of Indian and American (English) names and has emphasized the need of separate research work to be carried out on Indian names. We have identified two best features for classification namely 1, 2, 3-*grams-suffix* of a word & *isVowel* and shown that features like *isSonorant*, *numSyll* and *lenWord* are subsumed by the vowel ending feature. The best F_1-score reported by our system is 94.9% and we expect it to increase further as the training data increases. We hope that our results can be useful to the Indian NLP and ML community. Our training and test datasets would be made freely available for research purposes.

VIII. FUTURE WORK

The ratio of number of open syllables to the total number of syllables [8] in a name can be included as a feature for gender identification as females have a much higher corresponding ratio than males. Instead of taking *n-grams* of the whole word, first the word can be hyphenated into phonetic units and then their *n-grams* may be taken as a feature which would ensure a coherent classification of words having similar sounds. As vowel ending has been identified as an important and prominent feature in gender prediction, the training data can be partitioned into two sets, one having all the vowel ending words and the other containing the remainders. Then two different classifiers can be learnt from each of these two sets and one should be used to classify the vowel ending words and the other for the remaining ones. This partitioning of training data into two sets may ensure that all other features except vowel-ending are properly learnt by the classifier as well.

ACKNOWLEDGMENT

We would like to thank Mr. Gautam Kumar for his invaluable insights and suggestions. The graphs were plotted using gnuplot (www.gnuplot.info).

REFERENCES

[1] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.

[2] A. Bhargava and G. Kondrak, "Language identification of names with svms," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, June 2010, pp. 693–696.

[3] G.-F. Teng, W.-Q. Dong, J. Yang, and J.-B. Ma, "Gender identification for chinese e-mail documents," in *Proceedings of the Second International Conference on Innovative Computing, Informatio and Control*, ser. ICICIC '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 36–.

[4] H. cheng Lian, B. liang Lu, and S. Hosoi, "L.: Gender recognition using a min-max modular support vector machine," in *In: Proc. ICNC05-FSKD05, LNCS 3611*. Springer-Verlag, 2005, pp. 433–436.

[5] J. Yoo, D. Hwang, and M. S. Nixon, "Gender classification in human gait using support vector machine," *Lecture notes in computer science*, vol. 3708, p. 138, 2005.

[6] K.-H. LEE, S.-I. KANG, D.-H. KIM, and J.-H. CHANG, "A support vector machine-based gender identification using speech signal," 2008.

[7] R. E. Bellman, *Adaptive control processes - A guided tour*. Princeton, New Jersey, U.S.A.: Princeton University Press, 1961.

[8] A. S. Slater and S. Feinman, "Gender and the phonology of north american first names," *Sex Roles*, vol. 13, pp. 429–440, 1985, 10.1007/BF00287953.

[9] G. M., C. J., N. C., and T. N., "Vowel and consonant sonority and coda weight: A cross-linguistic study," in *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 2008.