Gender Prediction for Expert Finding Task

Daler Ali
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Nadeem Akhtar
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Hina Asmat
Dept. of Computer Science
Govt. S.E. College, Bahawalpur

Malik Muhammad Saad Missen
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Nadeem Salamat
Dept. of Basic Science and
Humanities
Khawaja Farid Univ. of Eng. And
Tech. RYK

Amnah Firdous
Dept. of Computer Science
COMSATS Institute of Information
Technology

Abstract—**Predicting gender by names is one of the most interesting problems in the domain of Information Retrieval and expert finding task. In this research paper, we propose a machine learning approach for gender prediction task. We propose a new feature, that is, *combination of letters* in names which gives 86.54% accuracy. Our data collection consists of 3000 Urdu language names written using English Alphabets. This technique can be used to extract names from email addresses and hence is also valid for emails. To the best of our knowledge, it is the first-ever attempt for predicting gender from Pakistani (Urdu) names written using English alphabets.**

Keywords—Urdu; Semantic Web; Gender Prediction; Expert Profiling; Machine Learning

I. Introduction

As internet becomes an intrinsic part of our lives, organizations tend to focus on automated solutions that can exploit the information available on the web. With the volume of increasing information on the web, the motivation for generating increased mass of knowledge is also increasing. However, it is very obvious that if technology is meant to bring benefits, it has to be able to support not only access to documented knowledge but also, most importantly, knowledge held by individuals [1]. To find and process such knowledge, expert finding task has been proposed by Information Retrieval research community.

The objective of an expert finding system is to help find people with the appropriate expertise through some intelligent automated techniques [2]. This task is very challenging because of rich set of information needs related to it. For example, finding the experts with particular set of skills within a particular domain or finding an expert from a specific geographical location. One of the most interesting and challenging tasks associated with expert finding task is gender prediction through expert's names. When searching for experts with the data available on the web, finding experts with a particular gender could be a very pertinent information need.

Gender prediction through names (or emails) is not only important for expert finding task only but also for many tasks like Co-reference Resolution, Machine Translation, Textual Entailment, Question Answering, Contextual Advertising and Information Extraction [3]. In literature, most of the work regarding gender prediction can be associated with author profiling tasks [4, 5] or gender prediction using names [3, 6] for expert finding tasks.

In this paper, we propose a machine learning approach to predict gender when written using English alphabets as these are mostly found on the web. We propose a feature (named as *combination of letters*) which gives better results when combined with existing proposed features (proposed for other languages). To the best of our knowledge, there is no previous work on this problem.

II. Related Work

Generally four types of work can be found when talking about gender prediction i.e.

- gender prediction using text,
- gender prediction using names,
- gender prediction using images,
- gender prediction using voice.

Gender predictions using images [7, 8] and voice [9, 10] are beyond scope of our work so we only discuss first two categories of work in this section.

A. Gender Prediction Using Text

Gender prediction using text is a sub-task of author profiling task. Author profiling, in general, is used to determine an author's gender, age, native language, personality type, etc [11]. It is a problem of growing importance in a variety of areas, including forensics, security and marketing. This is why it was also introduced as part of PAN (CLEF)[1] in year 2013 and continues to year 2016 as one of its core tasks. Gender prediction from text has been performed in several forms like blogs [12], electronic discourse [13], online social networks [14], and email [15].

[1] http://pan.webis.de/clef13/pan13-web/author-profiling.html

Researchers have been using style-based features (N-grams of POS tags in documents, punctuation symbols and number of *href* links [16, 17] etc.) as well as topic-based features for gender prediction from text (for example, males usually use words like 'daily life' to describe their work and whereas females use 'daily life' to describe their love or spiritual life).

B. Gender Prediction from names

Gender prediction from names is a challenging task; hence, one cannot find lot of work already done for this particular task. One of the foremost works done in this regard was on North American names [6]. In this work, researchers used morphological features of English language and find out many handful features of sound and language. Similarly, Tripathi and Faruqui [3] used support vector machine (SVM) approach for gender classification using Indian names. They used n-gram suffix along with other morphological features to classify males and females names.

C. What makes our Work Different

As discussed above, we could only find works on English and Indian names for gender prediction. Therefore, to the best of our knowledge there exists no work for Urdu names. Morphological analysis of American, Indian and Urdu names reveals their differences [3]. This makes our work different from the existing work.

Second thing that makes our approach distinct from existing approaches is the use of a new feature, that is, combination of letters.

Another difference is the size of data collection. We use around 3047 names (1729 female names while 1308 males) for our work while work for Indian names [3] used a collection of 2000 names (890 female and 1110 male names) while work on North American names [6] included 489 names (222 females and 267 male).

Last but not the least is the use of only textual features for identification of gender. We did not use sound-based features like syllables and sonorant consonant ending.

The following table shows feature-based analysis of our training data. We consider names long if they contain six or more letters.

TABLE I. TRAINING DATA FEATURE ANALYSIS

Feature	Male	Female
Length	36.03%	63.97 %
Vowel Ending	4.00%	49.77%

III. MACHINE LEARNING FEATURES

Previous works [3, 6] have used the following features differentiating between male and female names. We have only used a subset of the following features in our work because our focus is on using only text-based features.

Vowel Ending: Names of females generally end in a vowel while that of males in consonants.

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 [19]. 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, our data shows that females have longer names than males in Pakistani names when compared to Indian names where opposite trend has been reported [3].

A. Issues with Previously Used Features

Previously used features for gender prediction through names are a mixture of textual and speech based characteristics. However, we focus on using only textual based features which is more practical when predicting gender through names in real time. Each language has its own conventions inherited from the region where it is spoken. Therefore, textual features like vowel ending and length of the word might not behave the same way for Urdu language as for other languages. Therefore, we propose a new feature called "combination of letters" for gender prediction through Urdu language names. This feature tries to capture the consecutive or nonconsecutive combination of letters in names. "Combination of letters" could prove very useful when Urdu names (written in Roman Urdu) have the same ending letters and length because in that context these can't accurately distinguish between "Male" and "Female" names. Table 2 describes some examples of Urdu names (in Roman Urdu) with all three textual features.

TABLE II. DIFFERENT FEATURES EXAMPLES

Examples of Some Names with Features				
Name	Length	1gram	Combination of Letters	Gender
DANYAL	6	L	AN	M
FARYAL	6	L	AR	F

In this table, we can see that names "Danyal" and "Faryal" have same lengths and 1 gram ending but it is "combination of letters" which can help recognizing the gender of the name. To compute this feature, that is, "Combination of letter", we develop an algorithm which extracts this information automatically.

IV. EXPERIMENTS

A. Data Set

We prepared a dataset of Urdu (Pakistani) names ourselves from online web sites (containing Urdu names and their meanings) and old PTCL (Pakistan Telecommunication Limited) telephone directories available. All names are written in Roman Urdu script i.e. using English language alphabets. It is to be noted that most of the Urdu linguistics resources have been developed by Centre of Language Engineering[2] but we could not find a collection for Urdu names on their web site even they claim that they have developed one already [18].

[2] http://www.cle.org.pk/software/ling_resources.htm

Our data collection consists of 3047 Pakistani names. It consists of 1729 female while 1308 male names.

B. Classifiers

We use Decision Tree (J48), Support Vector Machine (SVM), K-nearest neighbor (Lazy-IBK) and Random Forest classifiers for individual as well as for different combination of features and compare their performances on results of testing data. We use 1828 (almost 60 percent of total data) name instances for building training model while rest of the 1219 instances are used as testing data. We use Weka[3] toolkit for our experimentation.

C. Results and Discussions

In this section, we describe the results obtained through different classifiers using individual or combination of different features.

a) Decision Tree: Following tables shows results for decision tree classifier.

TABLE III. RESULTS FOR DECISION TREE CLASSIFER

Classifier	Feature	Accuracy
J48	Length	61.53%
	Unigram	83.92%
	Combination of letters	57.91%
	2-gram	83.84%
	3-gram	78.261%
	Length and unigram	85.24%
	Length and combination of letters	61.03%
	Length and 2-gram	84.00%
	Length and 3-gram	78.67%
	Unigram and combination of letters	**86.55%**
	Length, unigram and combination of letters	84.74%
	2-gram, length and combination of letters	84.66%
	3-gram, length and combination of letters	79.33%
	2-gram and combination of letters	84.49%
	3-gram and combination of letters	78.75%

b) Support Vector Machine: Following tables shows results for SVM classifier.

TABLE IV. RESULTS FOR SVM CLASSIFER

Classifier	Feature	Accuracy
SMO	Length	57.01%
	Unigram	83.92%
	Combination of letters	57.91%
	2-gram	84.74 %
	3-gram	81.46%
	Length and unigram	84.74%
	Length and combination of letters	56.77%
	Length and 2-gram	84.74%
	Length and 3-gram	81.46%
	Unigram and combination of letters	**84.74%**
	Length, unigram and combination of letters	83.92%
	2-gram, length and combination of letters	84.74%
	3-gram, length and combination of letters	81.46%
	2-gram and combination of letters	84.74%
	3-gram and combination of letters	81.46%

c) K-Nearest Neighbour (KNN): Following tables shows results for KNN classifier.

TABLE V. RESULTS FOR KNN CLASSIFIER

Classifier	Feature	Accuracy
Lazy-IBK	Length	61.53%
	Unigram	83.92%
	Combination of letters	57.91%
	2-gram	83.98%
	3-gram	78.26%
	Length and unigram	84.90
	Length and combination of letters	61.61%
	Length and 2-gram	84.05 %
	Length and 3-gram	81.46%
	Unigram and combination of letters	**86.30%**
	Length, unigram and combination of letters	83.92%
	2-gram, length and combination of letters	85.15%
	3-gram, length and combination of letters	82.44%
	2-gram and combination of letters	84.90%
	3-gram and combination of letters	80.06%

d) Random Forest: Following tables shows results for random forest classifier.

TABLE VI. RESULTS FOR RANDOM FOREST CLASSIFIER

Classifier	Feature	Accuracy
Random Forest	Length	61.53%
	Unigram	83.92%
	Combination of letters	57.91%
	2-gram	83.92%
	3-gram	78.51%
	Length and unigram	84.66%
	Length and combination of letters	61.61%
	Length and 2-gram	83.75%
	Length and 3-gram	78.99%
	Unigram and combination of letters	**86.30%**
	Length, unigram and combination of letters	84.24%
	2-gram, length and combination of letters	85.23%
	3-gram, length and combination of letters	79.82%
	2-gram and combination of letters	85.23%
	3-gram and combination of letters	79.16%
	2-gram and combination of letters	85.06%
	3-gram and combination of letters	79.16%

Among individual features, unigram seems to be outperforming all other individual features for all classifiers. However, when unigram feature is combined with combination of letters, it further boosts up its performance to 86.55% from 83.92 % for decision tree, to 84.74 % from 83.92 % for SVM, to 86.30 % from 83.92% for KNN and Random Forest classifier. Accuracy of different classifiers for this combination is also shown in figure 1 (at end of the document). It is also very interesting to observe that 2-gram features seem to be playing more effective role than 3-gram features. We think it is because of the relatively shorter length of the names than other type of general words. Another

[3] www.cs.waikato.ac.nz/ml/**weka**/

positive aspect of these results is using 2-gram features with combination of letters which helps in further improving accuracy of 2-gram features.

V. Conclusions and Future Work

Results show that newly purposed feature. that is, finding combination of letters in name and unigram both together is the best feature for predicting gender by name written in URDU (Roman like English). We got highest accuracy of 86.54% with J48 classifier. It proves that using only textual based features can also improve gender prediction from names while existing works have achieved similar level of accuracy by using both i.e. speech and textual features.

While we have mentioned above that gender prediction from names is very important for expert finding task. We have focused on gender prediction in this work and we keep the task of predicting geographical background of the author from his /her text. For example, we might collect a data collection on same topic for authors from different locations and then find hidden patterns in their writings to determine their geographical background automatically. This task can be very helpful in determining political orientations or extremists attitudes for different topics.

References

[1] Yimam, Dawit. "Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR approach. ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop." (2000): 276-283.

[2] Karimzadehgan, Maryam, Ryen W. White, and Matthew Richardson. "Enhancing expert finding using organizational hierarchies." Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. 177-188.

[3] Tripathi, Anshuman, and Manaal Faruqui. "Gender prediction of Indian names." Students' Technology Symposium (Tec hSym), 2011 IEEE. IEEE, 2011.

[4] Argamon, Shlomo, et al. "Automatically profiling the author of an anonymous text." Communications of the ACM 52.2 (2009): 119-123.

[5] Estival, Dominique, et al. "Author profiling for English emails." Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07). 2007.

[6] 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

[7] Quanzeng You, Sumit Bhatia, Tong sum, Tiebo Luo," The eyes of the beholder: Gender prediction using images posted in Online Social Networks",2014 IEEE International Conference on Data Mining Workshop, Department of Computer Science, University of Rochester, NY

[8] Moghaddam, Baback, and Ming-Husan Yang. "Learning gender with support faces." Pattern Analysis and Machine Intelligence, IEEE Transactions on24.5 (2002): 707-711.

[9] Cornwall, Andrea. "Whose voices? Whose choices? Reflections on gender and participatory development." World development 31.8 (2003): 1325-1342.

[10] Wu, Ke, and Donald G. Childers. "Gender recognition from speech. Part I: Coarse analysis." The journal of the Acoustical society of America 90.4 (1991): 1828-1840.

[11] Rangel, Francisco, et al. "Overview of the author profiling task at pan 2013."CLEF Conference on Multilingual and Multimodal Information Access Evaluation. CELCT, 2013.

[12] Zhang, Cathy, and Pengyu Zhang. Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA, 2010.

[13] Thomson, Rob, and Tamar Murachver. "Predicting gender from electronic discourse." British Journal of Social Psychology 40.2 (2001): 193-208.

[14] Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh. "Predicting age and gender in online social networks." Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, 2011.

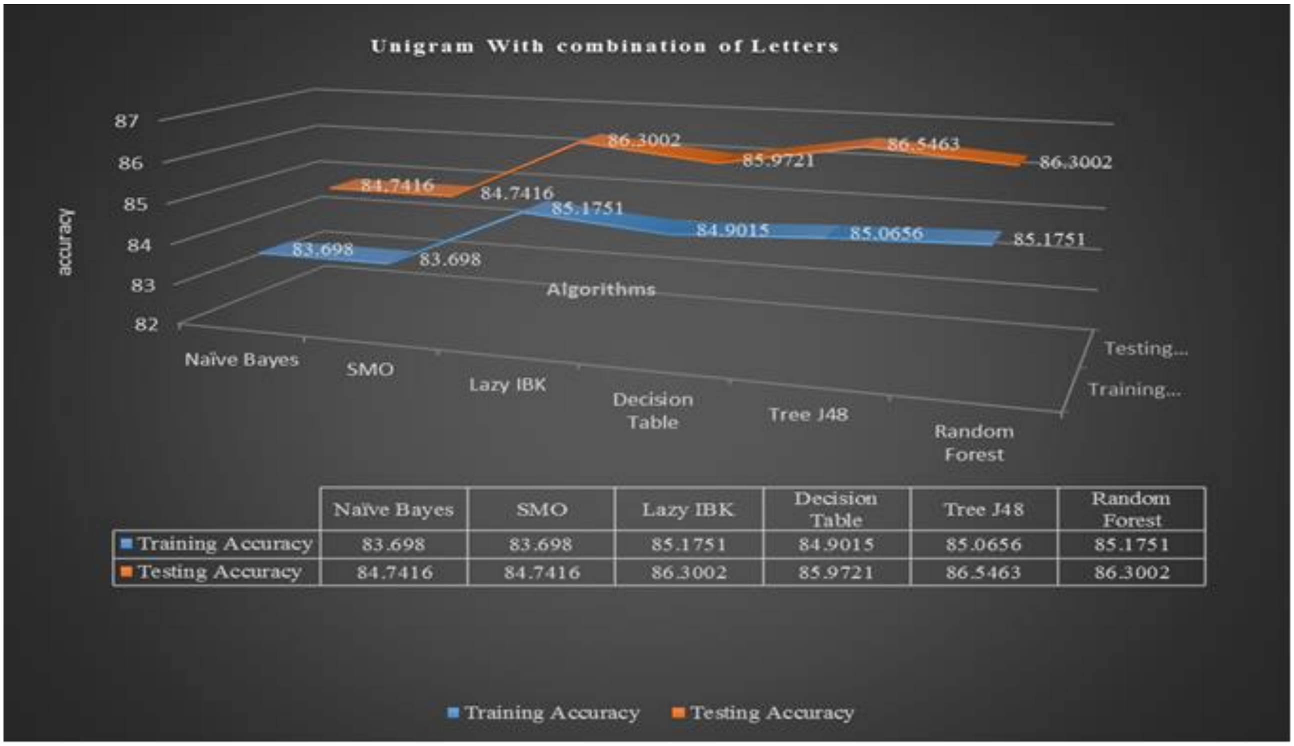
[15] Estival, Dominique, et al. "Author profiling for English emails." Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07). 2007.

[16] Argamon, Shlomo, et al. "Automatically profiling the author of an anonymous text." Communications of the ACM 52.2 (2009): 119-123.

[17] Schler, Jonathan, et al. "Effects of Age and Gender on Blogging." AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. Vol. 6. 2006..

[18] Hussain, Sarmad. "Resources for Urdu Language Processing." IJCNLP. 2008.

[19] Gordon, Matthew, et al. "Vowel and consonant sonority and coda weight: A cross-linguistic study." *WCCFL*. Vol. 26. 2008.



	Naïve Bayes	SMO	Lazy IBK	Decision Table	Tree J48	Random Forest
■ Training Accuracy	83.698	83.698	85.1751	84.9015	85.0656	85.1751
■ Testing Accuracy	84.7416	84.7416	86.3002	85.9721	86.5463	86.3002

Fig. 1. Result