Feature Selection Techniques for Gender Prediction from Blogs

Shahana P.H
Department of Computer Science
and Engineering
SCMS School of Engineering
and Technology
Ernakulam,Kerala
shahana2702@gmail.com

Bini Omman
Department of Computer Science
and Engineering
SCMS School of Engineering
and Technology
Ernakulam,Kerala
binireni@gmail.com

***Abstract*—The goal of this paper is to identify gender of blog authors. Features such as POS tags, unigram (words+ punctuations), bigrams and word classes are considered. To synthesis/rank features we are using Mutual information, Chi-square and Information gain methods. The dataset is the collection of 3227 blogs originally derived from blogs set, and among them 1679 were written by male and 1548 were written by female. The results were obtained using 10-cross fold validation. Unigram of words gave better accuracy of 78.81% in comparison with the other features. We found that chi-square is the best in ranking features. The classification is done using Multinomial Naïve Bayes Classifier, and different kernel functions of SVM such as PolyKernel, Puk, Normalized PolyKernel and RBFkernel.**

***Keywords*—Classification, feature selection, Multinomial Naïve Bayes, SVM classifier, WEKA classifier.**

I. Introduction

Blogs are written by individuals, each having their own writing styles. Determining the gender of a blog's author is a complex problem. Classification of gender of blog authors will give hidden information about their like and dislikes. The prediction of gender of blog authors by evaluating their writing styles has an importance in many domains. The motivations of this paper are (1) Sentiment analysis: knowing people's (Men and Women) opinion on products and services (2) Targeted Advertising: Knowing gender information with significant accuracy can help targeted advertising which will boost sales. The preprocessing stage of gender classification is the most challenging phase since the blog data contains elongated words, wrong spelling words, smiley etc. We are analyzing the relationship between the gender and writing styles of authors of the blog. The main objective of this paper is to predict the gender of blog author. Also we are interested in knowing the effective feature that can provide the better result as well as the best feature selection method. We have considered unigram of words, word classes, POS tags and bigram of words.

Remaining sections are organized as follows: Section (II) delivers some related works on the domain of gender classification of blog authors; Section (III) contributes an overview of our proposed methodology; Section (IV) gives experiment and result; Section (V) provides inference. And paper ends with the conclusion and future work in section (VI).

II. Related Work

Schler et al. [1] used style-related and content-related features. Style-related features include relevant parts-of-speech, function words, blog specific features and hyperlinks. Content-related features consist of simple content words and special classes of words from the handcrafted LIWC categories. The stylistic and content features together offered better classification accuracy. The corpus used for this experiment contained, blogs those were collected from blog sites like blogger.com in August 2004. In total corpus 71,000 blogs were included. The accuracy of this experiment was 80.1%.

In this paper, Argamon et al. [2] was used EG algorithm to extract the relevant features for the classification of documents. The feature list included large number of determiners and POS as male indicators and pronouns as female indicators. 604 documents were collected from the British National Corpus (BNC) to create corpus for this experiment. Each sample in the corpus was labeled with gender. All words in the document were POS tagged using the BNCs 76 tags. This experiment has acquired an accuracy of about 80%.

Nowson et al. [4] compared the genre of documents collected from the British National Corpus (BNC). The corpus consists of more than 4000 files, with 100 million words of both spoken and written English. In this experiment authors calculated F-score of each sample in the corpus. F-score measures contextually formality of the sample. The F-score can differentiate the genders of writers.

Koppel et al.[5] proved that the combination of simple lexical and syntactic features automated text categorization techniques can predict the authors gender with approximately

978-1-4799-3486-7/14/$31.00

80% accuracy. They have used BNC dataset which contains 920 documents in British English with the class of the authors gender. The feature sets used for this study composed of function words only (FW), parts-of-speech only (POS) and function words and parts-of-speech (FWPOS) together.

Houvardas and Stamatatos et al. [6] used Information gain (IG) as baseline. The Reuters Corpus Volume 1 (RCV1) for the English language including over 800,000 newswire stories was the corpus used for this experiment. In order to capture stylistic feature for news articles in Reuters corpus, authors applied N-grams of characters.

Ansari et al. [7] performed their experiment with 3 features such as frequency counter, Term Frequency-Inverse Document Frequency(TF-*idf*), POS tags. Frequency counter is the ratio of the count of the occurrence of a token in a given sample to the total number of words in that particular sample. The second feature calculated the relevance of each term. This machine learning experiment used 100 blogs and the total number of words was 4523. The classifier used for the classification was Naïve Bayes. The authors achieved the best prediction accuracy as 67.23%. They have used ZeroR as the classifier.

Mukherjee et al. [8] performed their experiment on real-life blogs collected from blog sites. They have used F-measure, Factor analyses and Word classes, N-grams of POS tags, gender preferential features, and POS sequence patterns as features. The authors used SVM regression, SVMLight and Naïve Bayes for creating classification model. Also they were used ensemble feature selection methods for ranking feature set. SVM_regression performed well with ensemble feature selection and POS sequence patterns. They achieved 88.56% accuracy.

Koppel et al. [10] discovered that the gender of author can be identified by analyzing the writing style of authors. In this paper, they employed machine learning algorithms on a gender labeled corpus of 566 documents collected from the British National Corpus (BNC). The feature set is composed of 405 function words and a list of N-grams of parts-of-speech extracted using the BNC's tag set of 76 parts of speech and punctuation marks. The results were obtained using 10 fold cross-validations. The feature set consisting of function words (FW), parts-of-speech (POS) and combination of function words and parts-of-speech (FWPOS). While using function words for classification, 73.7% of the documents were correctly classified, by using parts-of-speech only about 70.5% documents were classified, and for the ensemble feature set(FWPOS),accuracy was 77.3%.

III. Proposed methodology

This section explains the proposed method which is applied on the dataset. The method includes three main steps: (1) Preprocessing (2) Feature selection (3) Classification.

A. Dataset

The corpus[15] contains 3227 samples originally derived from blogs. Among 3227 samples, 1679 were written by male and 1548 were written by females. The average length of blog written by male is 250 words and 330 words for female.

B. Preprocessing

The samples in the dataset should be preprocessed before performing any type of operation in it, since it contains some Non-ASCII characters, slang words etc. The preprocessing includes

1) Upper to lower case conversion: For the easiness of feature selection all the data should be converted into lower cases.

2) Normalization: All words with apostrophizes should be replace with its original form. E.g. don't → do not.

3) Non ASCII removal: All non ASCII characters are removed from the samples.

4) Remove new lines: The dataset contains some unwanted new lines that are also removed before the feature selection phase.

5) Elongated characters replacements: Some words in the dataset contain characters more than the normal count. E.g. beesstt→best, aaahh→ah. These types of words should also replace with its normal form.

6) Stopword removal: Stop words in the English language are a, an, the, is. So we have removed all the words whose length is less than 3.

7) Stemming: All words should be transferred to its root form. E.g.: Activity, Activated, Activities→Active. For stemming our words we have used the porter stemmer [11].

C. Features

We are considering different classes of features in this paper. Those are listed below.

1) Unigrams: The main features used in this paper are words and punctuatoins that people used to write the blogs. There are 49042, which contains words as well as punctuations. All words are in the stemmed version.

2) POS tags: POS tags are used to capture the heavy stylistic and syntactic information. All the samples are tagged using part of speech (POS) tagger [13]. Since the tagger provides 36 tagger, we are only using 6 popular tags among them, such as NN, JJ, VB, PRP, UH, RB. We are taking average of all these tags in a sentence rather than presence or absence of these tags in a sample.

3) Word classes: There are some words that are coming under the same group. Those words are mapped to the common class label. This process is called meaning extraction methods. We have considered twenty classes[9]. There will be 1000 words in these 20 classes. Each class contains 323 function words as well as 677 different content words.

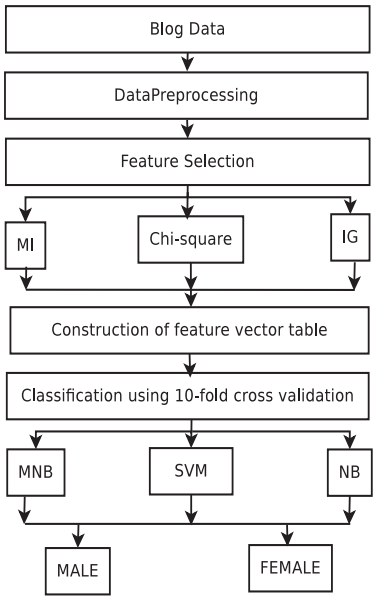

Fig. 1. Architecture of Proposed Methodology

4) Bigrams of words: 58000 bigrams are extracted from the corpus, which contains words as well as punctuations. All words are in the stemmed version. And we are taking presence or absence rather than frequency of bigrams.

D. Feature Selection Methods

We are having a large number of features, among them some features are contributing more for classification. In order to select subset of relevant features we used the following feature selection methods.

1) Mutual Information (MI): MI term selects features that are not uniformly distributed among the gender classes because they are informative of their classes. And we can see that MI giving more importance to the rare term.

$$MI(f,c) = \sum_{c \in C} \sum_{f} P(f,c) log \frac{P(f,c)}{P(f)P(c)} \qquad (1)$$

Where *P(f,c)* indicates the joint probability distribution function, *P(f)* and *P(c)* denotes the marginal probability distributions of f and c, and C is the classes: Male and Female.

2) Information gain: Information gain is the most commonly used feature selection method in the field of machine learning. It calculates the relevance of a feature for prediction of gender class by analyzing the presence or absence of a feature in a document.

$$IG(f) = -\sum_{c,\bar{c}} P(c) log(c) + \sum_{f,\bar{f}} P(f) \sum_{c,\bar{c}} P(c|f) log P(c|f) \qquad (2)$$

3) Chi-square: χ^2measures how much expected counts and observed counts deviate from each other.

$$\chi^2(f,c) = \frac{N(WZ - YX)^2}{(W+Y)(X+Z)(W+X)(Y+Z)} \qquad (3)$$

W, X, Y, Z denotes the frequencies, indicates the presence or absence of feature in the sample. *W* is the count of samples in which feature f and c occurred together. And by using the TABLE 1 we can find what each symbol indicates. *N=W+X+Y+Z.* And f is the feature and c is the class.

TABLE 1
2×2 CONTINGENCY TABLE OF FEATURE (f) AND CLASS(c)

	c	$\bar{c}$
f	W	X
$\bar{f}$	Y	Z

IV. Experiment and Result

A. Experimental setup

The experiments are performed on the blog data. Before performing the classification of gender, we apply preprocessing step in the dataset [15]. The features are selected after preprocessing of the samples in corpus. The preprocessing contains stemming; it is applied to reduce the inflected forms of words to their root form. This is done by the Porter Stemmer module which uses the Porter stemming algorithm [12].

For feature selection, we will use the selection criteria, which are Mutual information, Information gain and Chi-square. The score is calculated for each feature with respect to the two classes, which are then sorted in decreasing order of their score. In order to find the optimal feature length,the classification is done on features of different length, extracted from these sorted list. The MNB and LibSVM, are used to prepare the classification model. The MNB and LibSVM with kernel functions [9] such as linear kernel, RBF kernel (g=0.5), polynomial (d=1) implementation of WEKA [14] is used in this work.

B. Result

Unigram of words, Word classes, POS tags and Bigram of words are used as features. Bigrams and POS tags are not producing good result, because these features are not enough to differentiate the classes. When we are using POS tags, we are getting generalized form of feature. All features

are mapped to its parts of speech form, for example if male feature (verb) is mapped to VB tag and female feature (verb) is also mapped to VB tag so that POStags could not perform well. We are also considering Bigram of words, Since it does not produce good results, we are not including that result in this paper. Unigram provides better results compared to the other features. The features are ordered using scores of mutual information, chi-square and information gain. The accuracy of the classifier is evaluated for different feature sets of varying length to determine optimality of the feature set. A feature set those results in higher accuracy is considered to be candidate for the precise model. For unigram features (bag-of-words), the classifier is tested on feature sets of size 100-10,000. Also we found to produce higher accuracy in all cases (refer to the results tabulated as under).

TABLE 2 depicts the percentage accuracy of unigram features, using MI feature selection criteria. Multinomial Naïve Bayes classifier gives better accuracy of 72.35% for 2000 features. TABLE 3 describes the percentage accuracy of unigram features, using IG feature selection criteria. Multinomial Naïve Bayes classifier gives better accuracy of 74.09% for 10,000 features. TABLE 4 depicts the percentage accuracy of unigram features, using Chi-square feature selection criteria. Multinomial Naïve Bayes classifier gives better accuracy of 78.81% for 10,000 features.

Compared to MI,IG feature selection methods,**Chi-square** gives better result of **78.81% accuracy**. We also consider different classifiers such as MNB and LibSVM with kernel functions [9] such as linear kernel, RBF kernel (g=0.5), polynomial (d=1).Among these classifiers **Multinomial Naïve Bayes classifier** performs well.

TABLE 2
PERCENTAGE ACCURACIES OF UNIGRAM FEATURES FOR MI USING DIFFERENT CLASSIFIERS

FL	MNB	linear kernel	RBF kernel g=0.5	Polynomial d=1
200	68.29	66.99	64.27	62.34
500	70.8	67.43	63.40	58.90
1000	72.07	67.21	57.60	55.12
2000	72.35	68.32	54.38	53.45
5000	71.95	67.33	52.77	52.02
10,000	69.32	65.6	52.71	52.02

TABLE 3
PERCENTAGE ACCURACIES OF UNIGRAM FEATURES FOR IG USING DIFFERENT CLASSIFIERS

FL	MNB	linear kernel	RBF kernel g=0.5	Polynomial d=1
200	52.24	52.52	52.71	52.02
500	53.73	54.32	53.82	52.02
1000	55.40	55.90	54.91	52.02
2000	58.47	58.47	57.42	52.02
5000	68.44	64.77	60.12	52.02
10,000	74.09	65.60	59.71	52.02

TABLE 4
PERCENTAGE ACCURACIES OF UNIGRAM FEATURES FOR χ^2 USING DIFFERENT CLASSIFIERS

FL	MNB	linear kernel	RBF kernel g=0.5	Polynomial d=1
200	69.91	68.42	64.67	64.67
500	73.31	69.22	59.71	62.44
1000	74.83	69.35	54.01	57.79
2000	76.38	71.21	52.64	54.83
5000	77.71	70.31	52.71	52.27
10,000	78.81	69.72	52.71	52.02

TABLE 5
PERCENTAGE ACCURACIES OF FEATURES

Features	Accuracy (%)
Unigram Of Words + Word Classes	78.81
POS-tags	56.52

In the prior work [11] authors have got **72.01%** accuracy for information gain. They were considered Unigrams,POStags and Word factor analysis and Sentence length of each samples as feature. But in our method we have got **78.81%** accuracy for Chi-square feature selection method,using Unigrams,Bigrams,POStags and Word classes.

V. CONCLUSION AND FUTUREWORK

This paper suggests the problem of gender classification. We found that unigram is the best feature for the gender classification. Chi-square feature selection methods gave better accuracy compared to the MI, IG. Multinomial Naïve Bayes classifier produced good classification result of 78.81%. This study was conducted on a real-life blog data set. It was observed that when number of features selected is small the classification accuracy is low. The accuracy increases when the number of selected features increases. As a future work we can suggest that ensemble feature selection technique, it would be useful to perform additional experiment on this work.

REFERENCES

[1] Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W, *"Effects of Age and Gender on Blogging"*, In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (Vol. 6, pp. 199-205).

[2] Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). *"Gender, genre, and writing style in formal written texts"*. TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-, 23(3), 321-346.

[3] Yan, Xiang, and Ling Yan.*"Gender Classification of Weblog Authors"*.AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs".2006.

[4] Nowson, Scott, Jon Oberlander, and Alastair J. Gill.*"Weblogs, genres and in-dividual differences."*Proceedings of the 27th Annual Conference of the Cog-nitive Science Society. Vol. 1666.2005.

[5] Koppel, M., Argamon, S.,Shimoni, A. R. (2001).*"Automatically determining the gender of a texts author"*. Bar-Ilan University Technical Report BIU-TR-01-32.

[6] Houvardas, John, and Efstathios Stamatatos.*"N-gram feature selection for authorship identification."*Artificial Intelligence: Methodology, Systems, and Applications. Springer Berlin Heidelberg, 2006. 77-86.

[7] Ansari, Yasir Zafar, Shams Abubakar Azad, And Halima Akhtar.*"Gender Classification Of Blog Authors."*

[8] Mukherjee, Arjun, and Bing Liu. *"Improving gender classification of blog authors."* Proceedings of the 2010 conference on Empirical Methods in natural Language Processing. Association for Computational Linguistics, 2010.
[9] Ben-Hur, Asa, and Jason Weston. *"A user's guide to support vector machines."* Data mining techniques for the life sciences. Humana Press, 2010. 223-239.
[10] Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. *"Automatically categorizing written texts by author gender."* Literary and Linguistic Computing17.4 (2002): 401-412.
[11] Zhang, Cathy, and Pengyu Zhang. *"Predicting gender from blog posts"*. Technical Report. University of Massachusetts Amherst, USA, 2010.
[12] http://tartarus.org/~martin/PorterStemmer/python.txt
[13] http://pypi.python.org/pypi/topia.termextract
[14] http://www.cs.waikato.ac.nz/ml/weka/
[15] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html