Exploiting Wikipedia Categorization for Predicting Age and Gender of Blog Authors

K Santosh, Aditya Joshi, Manish Gupta, Vasudeva Varma

International Institute of Information Technology
Hyderabad, India

Abstract. For privacy reasons, personally identifiable information like age and gender of people is not available publicly. However accurate prediction of such information has important applications in the fields of advertising, forensics and business intelligence. Existing methods for this problem have focused on classifier learning using content based features like word n-grams and style based features like Part of Speech (POS) n-grams. Two major drawbacks of previous approaches are: (1) they do not consider the semantic relation between words, and (2) they do not handle polysemy. We propose a novel method to address these drawbacks by representing the document using Wikipedia concepts and category information. Experimental results show that classifiers learned using such features along with previously used features help us achieve significantly better accuracy compared to the state-of-the-art methods. Indeed, feature selection shows that our novel features are more effective than previously used content based features.

1 Introduction

In recent years, exponential increase in textual information has sparked interest in automatically predicting personally identifiable information (PII) such as age and gender of users. Automatic prediction of age and gender has various applications in the fields of forensics, business intelligence and security.

Research in identifying author's age and gender started with extensions of the earlier works on categorization and classification of text. Koppel et al. [2] exploited combinations of lexical and syntactic features to infer the gender. Koppel et al. [3] explored differences in writing style and content between male and female bloggers as well as among authors of different ages. Meina et al. [4] used an ensemble based classification method to determine age and gender. They used various content and style based features. The overview paper of the PAN Author Profiling task [5] discusses various approaches used by their participants. It states that participants used content based (bag of words, word n-grams, slang words, etc.), style based (POS, readability measures, punctuations etc.) features, and that the ensemble of all features performed better. However, the two major issues with the content based features used in above works are: (1) they do not consider the semantic relation between words, and (2) they do not handle polysemy. Our method addresses these issues by representing a document in the feature space of Wikipedia concepts and categories.

2 The Proposed Approach

People of different gender and age have different interests. Hence, there is a lot of contextual difference between blogs written by different people. In our approach we explore these contextual differences to predict age and gender of an author of a text. Our approach consists of two phases: Semantic representation of documents, and age and gender prediction.

2.1 Semantic Representation of Documents

We extracted Wikipedia concepts related to the entity mentions in the text for each document in the training corpus. For every Wikipedia concept, we found its categories in Wikipedia. In order to get an exhaustive list of categories, we recursively collected the categories up to five levels. We refer to the list of categories at level i as $Cat\ L_i$. Our final document representation thus consists of a collection of Wikipedia concepts and categories. We refer to these features as Wikipedia semantic features.

Preprocessing Data: The text from blogs is preprocessed to remove HTML tags and unwanted boilerplate content like advertisements to get the clean data.

Entity Linking: We used TAGME API [1] to find Wikipedia concepts in the text. TAGME uses anchor text found in Wikipedia as spots (sequence of terms which are ambiguous) and the pages linked to them in Wikipedia as their possible senses. TAGME tackles the ambiguity and polysemy problems in the potentially many available anchor-page mappings by finding the collective agreement among them via scoring functions which are both fast to compute and effective.

Finding Parent Categories for Wikipedia Concepts: For all the Wikipedia concepts extracted in the previous step, their parent categories up to five levels are extracted. We created a Wikipedia category network using Wikipedia's category corpus and the networkx library[1] and traversed up to five levels on this network to obtain all parent categories. Semantically related words get mapped to similar set of Wikipedia categories at various levels; thus semantic relations between the words get captured in our approach.

2.2 Age and Gender Prediction

To predict the author's profile, i.e., age group and gender of the author of a document, we used two machine learning classification models namely, K-Nearest Neighbors (KNN) and Support Vector Machines (SVM).

KNN: Given a test document q, we represented the documents in terms of Wikipedia semantic features as mentioned in Subsection 2.1. We used Okapi-BM25F [6] distance metric to compute k nearest neighbors to the test document. While computing Okapi-BM25F, we considered Wikipedia concepts and category at different levels as the fields.

SVM: We also learned SVM classifiers for age and gender prediction using Wikipedia semantic features, and content and style features.

[1] http://networkx.github.io/

3 Experiments

Dataset: We used PANTraining and PANTesting datasets provided by the 2013 PAN Author Profiling Task[2]. The class labels are Male and Female for gender, and three groups of age (10s: 13-17 yrs, 20s: 23-27 yrs and 30s: 33-47yrs). Dataset details are shown in Table 1. We divided PANTraining into two parts: 70% for training, 30% for validation.

KNN Classifier: For KNN classifier, we learned the boost factor for each field c using the validation set as $boost_c = \frac{AccWith_c}{AccWithout_c}$ where, $AccWith_c$ is the accuracy obtained by using field c alone, and $AccWithout_c$ is the accuracy obtained by using all the other fields except c. Figures 1 and 2 show that each of the features are important for the prediction task.

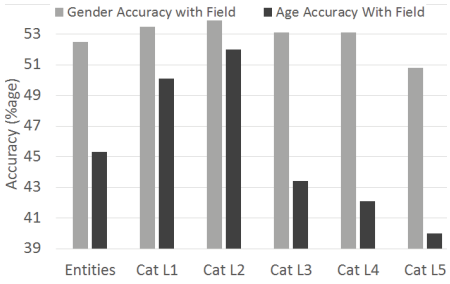

Fig. 1. Accuracy with Particular Field Considered

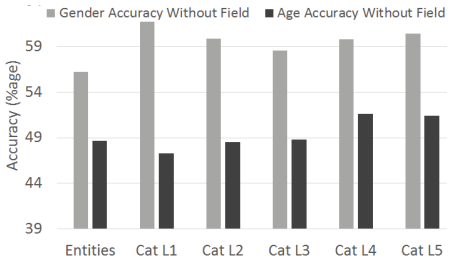

Fig. 2. Accuracy with Particular Field Ignored

On validation data, we obtained best accuracy at k=5 for gender prediction and k=7 for age prediction. Hence, we use these values of k while testing.

SVM: For SVM, along with Wikipedia semantic features, the following features are also used. (a) **Content based features**: These features analyse the content of the blogs. Koppel et al. [2] used unigrams as content features. In this work, we use unigrams, bigrams and trigrams as content based features. (b) **Style Features**: These are features which capture people's writing styles. In this work, we use POS n-grams (upto trigrams) as style features. Various combination of above mentioned features are used for building classifiers. The size of the feature vector for these feature sets are listed in Table 2.

Age	#Train Instances	#Test Instances
10s	17200	1776
20s	85800	9174
30s	133600	14408

Table 1. Dataset Details (Equal Distribution for Males and Females)

	Gender	Age
word n-grams	50000	61781
POS n-grams	16000	18000
Wikipedia Semantic	300000	306910

Table 2. Number of Features for Gender and Age Classifiers

We learned the parameters of SVM using 10-fold cross validation on PANTraining data. Our experiments did not find other kernels to perform any better than the linear kernel. Table 3 compares accuracies of approaches using different combination of word n-grams, POS n-grams, our Wikipedia Semantic features and state-of-the-art method

[2] http://www.webis.de/research/corpora/corpus-pan-labs-09-today/pan-13/pan13-data/

on the PANTesting data. We also compared with Meina et al. [4]'s method that obtained the best accuracy in the PAN Author Profiling Task at CLEF 2013.

Features	Classifier	Gender	Age
Wikipedia semantic	KNN	56.42	61.38
Wikipedia semantic	SVM	56.61	61.85
Word n-grams	SVM	53.21	56.79
POS n-grams	SVM	54.56	57.37
Wikipedia semantic + word n-grams	SVM	57.27	62.67
Wikipedia semantic + POS n-grams	SVM	58.39	63.29
Wikipedia semantic + word n-grams + POS n-grams	SVM	**62.12**	**66.51**
Meina et al. [4]		59.21	64.91

Table 3. Accuracy Comparison of Various Approaches on PANTesting Data

Accuracy comparisons in Table 3 show that our Wikipedia semantic features are better than the word n-gram based features, and combination of all features yields the best accuracy.

4 Conclusions

We studied the problem of age and gender prediction. We leveraged the document representation using Wikipedia concepts and category information as features for KNN and SVM classification. Experimental results show that the proposed approach beats the best approach for a similar task at CLEF 2013. By enhancing the entity linking part of the proposed system, overall accuracy of the age and gender prediction can be further improved. In the future, we would like to limit our reliance on entity linking and also explore other learning algorithms and robust features that can help in predicting the age and gender of the author of a document.

References

1. P. Ferragina and U. Scaiella. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *Proceedings of the* 19^{th} *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628. ACM, 2010.
2. M. Koppel, S. Argamon, and A. R. Shimoni. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17:401–412, 2003.
3. M. Koppel, J. Schler, S. Argamon, and J. Pennebaker. Effects of Age and Gender on Blogging. In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.
4. M. Meina, K. Brodzinska, B. Celmer, M. Czokow, M. Patera, J. Pezacki, and M. Wilk. Ensemble-based Classification for Author Profiling Using Various Features. *The Notebook for 2013 PAN at the Conference and Labs of the Evaluation Forum (CLEF)*, 2013.
5. F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the Author Profiling Task at PAN 2013. *The Notebook Papers of the Conference and Labs of the Evaluation Forum (CLEF) 2013 LABs and Workshops*, pages 23–26, 2013.
6. H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard Tracks. In *Proceedings of the 2004 Text REtrieval Conference (TREC)*, 2004.