Author Profiling from Personal Content Blogs

Aayush Singhal
12CS10002

Aseem Patni
12CS10008

Soham Dan
12CS10059

Bhushan Kulkarni
12CS30016

Pranay Yadav
12CS30025

Shubham Saxena
12CS30032

Sruthi Warrier
Mentor

Anurag Verma
Mentor

Suman Kalyan Maity
Mentor

ABSTRACT

This project aims to predicting personally identifiable information (PII), such as age and gender of the author by extracting features from his/her personal content blog texts. We intend to define the state-of-the-art in the field and overcome the shortcomings of the prior works in the personality recognition tasks. This report describes our progress so far and contains details about our future work-flow.

Keywords

Authorship Profiling, PII, Blogosphere

1. INTRODUCTION

Though the enormous impact of social media on our daily life, we observe a lack of information about those who create the contents. In this regard, author profiling tries to determine the gender, age, native language or personality type of authors by analyzing their published texts. In this study, we focus on building a system to identify only the gender and age of the authors. Other authorship details will be a part of the future work in this area. Author profiling is of growing importance: E.g., from a marketing viewpoint, companies may be interested in knowing the demographics of their target group in order to achieve a better market segmentation; from a forensic viewpoint, determining the linguistic profile of a person who wrote a ”suspicious text” may provide valuable background information.

This study is targeted towards partial fulfillment of requirements for *CS60057: Speech & Natural Language Processing* during Fall 2015, under the guidance of Prof. Pawan Goyal.

The remainder of this paper is organized as follows. Section 2 describes the corpus, Section 3 covers the proposed approach, Section 4 presents the results obtained so far, Section 5 discusses the evaluation measures, Section 6 contains details about the future work-plan and Section 7 concludes by listing the work done by the individual team-mates.

2. DATA-SET

2.1 Corpus

We have used the following two corpora for this study:

- Blog Authorship Corpus
  The Blog Authorship Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words or approximately 35 posts and 7250 words per person. All bloggers included in the corpus fall into one of three age groups — ”10s” [13-17], ”20s” [23-27], ”30s” [33-47]. For each age group there are an equal number of male and female bloggers.
- PAN’14 Corpus
  As a part of the Author Profiling Shared Task in PAN ’14, this corpus was made available for use during the competition. This data-set originally consists of blog posts, tweets and social media texts written in both English and Spanish as well as hotel reviews in English. We have considered only the subset which contains blog posts. All bloggers included in the corpus fall into one of these age groups: [18-24], [25-34], [35-49], [50-64], [65-xx]. The corpus incorporates a total of 2278 posts, 148 authors or on an average 15 blogs per author.

We split ourselves into two groups, one for extracting features from the Blog Authorship Corpus (Pranay, Shubham & Soham) and the other from the PAN ’14 Corpus (Aayush, Aseem & Bhushan).

2.2 Data Cleaning & Extraction

- Blog Authorship Corpus
  The corpus contains 19,320 XML files, each pertaining to a particular author, identified by the unique filenames. Each XML file contains date when the blog was posted followed by the post itself. All the HTML links in the post are replaced by a unique tag *’urlLink’* to mark their presence. We cleaned the data by discarding empty blog posts and ignoring posts which contain only HTML links and no text. We then exported this refined data to a JSON file, on which further analysis will be carried out.
- PAN ’14 Corpus
  This corpus contains 148 XML files, each pertaining

to a particular author. Each XML file contains the Author’s unique ID and blogs written by the Author. The blog text is present in `CDATA` section. To parse this text, we wrote a regular expression to remove the HTML tags, translated HTML entities like `’&amp;’`, `’&ldquo;’` to their usual textual counter-parts like ’&’,’”’. We then dumped this data as to a JSON file, on which further analysis will be carried out.

3. APPROACH

After obtaining the JSON files containing refined data from the both the corpora, we now start extracting features from this data-set. We shall be focusing our attention towards building two kinds of classifiers — 1) Binary classifier for classification of gender and 2) Multi-label classifier for classification of age into predefined class labels. Later, we will also consider the possibility of predicting the age by fitting regression models, by working under the assumptions that — 1) There are enough data points for the model to fit accurately and 2) Age behaves like a continuous variable.

3.1 Exhaustive Feature Set

In this study, we shall consider the following types of features for building our classifiers:

- Content-based Features viz. # of HTML links in the blog, # of named entities used, # of non-word errors, # of discourse relations within the text, # of quotations used in the text, # of references to past or future within the text, # of facts & figures used, # of times opinions are expressed, overall sentiment score of the blog, # of words having character flooding (like ’helloooo’).

- Style-based Features viz. distribution of POS tags, distribution of punctuation tags, readability measure of the blog (SMOG & Fischer), # of co-references (usage of pronouns), average sentence length, usage of figures of speech by the author (like metaphor, alliteration).

- Semantic Features: Latent Semantic Analysis (LSA) of the blogs to identify set of topics authors, belonging to a particular gender or age group, blog about.

These features will be extracted for both the corpora.

3.2 Extracting Feature Vectors

Methodology to extract feature vectors from the Blog Authorship Corpus is described below:

- Content-based Features:
  1) # of HTML links: count the occurrence of *’urlLink’* in the body of the blog.
  2) # of named entities: Using Stanford NLTK APIs to tokenize blog text into sentences, perform POS Tagging and then extract named entities (NE) from the tagged sentences.
  3) # of non-word errors: Using Stanford NLTK’s word corpus `nltk.corpus.words.words()` to keep a count of non-word errors in the blog.
  4) # of discourse relations within the text: Using a Java-based end-to-end PDTB-styled Discourse Parser to identify implicit & explicit discourse relations and keeping a count of each of them.
  5) # of quotations used in the text: Checking for occurrences of ’”’ in the blog text.
  6) # of references to past or future: Checking for occurrences of the following set of words and phrases - [’years ago’,’years from now’,’in the past’,’in future’,’once upon a time’,’^\d{4}$’]. The last entry is a regex for detecting reference to a year.
  7) # of facts & figures used: To be decided.
  8) # of times opinions are expressed:

3.3 Feature Subset Selection (FS)

Dimensionality reduction (DR) and Feature Subset Selection (FS) are two techniques for reducing the attribute space of a feature set. The main idea of FS is to remove redundant or irrelevant features from the data set as they can lead to a reduction of the classification accuracy and to an unnecessary increase of computational cost. The advantage of FS is that no information about the importance of single features is lost. For now, we will focus our attention on using FS over DR because DR can decrease the size of the attribute space strikingly. Another important disadvantage of DR is the fact that the linear combinations of the original features are usually not interpretable and the information about how much an original attribute contributes is often lost. If possible, we might also try using DR technique (PCA) to our feature set and evaluate the improvement in the accuracy of the resulting classifier, if any.
There are three types of feature subset selection approaches:

- Filters:
  Filters are classifier agnostic pre-selection methods which are independent of the later applied machine learning algorithm. . Besides some statistical filtering methods like Fisher score or Pearson correlation, information gain is often used to find out how well each single feature separates the given data set.
  The overall entropy I of a given dataset S is defined as:

$$I(S) := -\sum_{i=1}^{C} p_i log_2 p_i \tag{1}$$

  where C denotes the total number of classes and p_i the portion of instances that belong to class i. The reduction in entropy or the information gain is computed for each attribute according to:

$$IG(S,A) = I(S) - \sum_{v \in A} \frac{|S_{A,v}|}{|S|} I(S_{A,v}) \tag{2}$$

  where v is a value of A and $S_{A,v}$ is the set of instances where A has value v.

- Wrappers:
  Wrappers are feedback methods which incorporate the ML algorithm in the FS process, i.e. they rely on the performance of a specific classifier to evaluate the quality of a set of features. Wrapper methods search through the space of feature subsets and calculate the estimated accuracy of a single learning algorithm for each feature that can be added to or removed from the feature subset.

We shall focus our attention towards using Filters to perform FS in this study.