Author Age Attribution in Blogs using Syntactic Stylometry

Anthony Kunnel Jose(108817537) Jagat Sastry (108721027) Rucha Mahadik (108672217)

Hypothesis

Bloggers of a particular age range follow a certain syntactical structure in their blog posts, which can be mined using stylometric techniques based on PCFG parse trees.

Motivation

Predicting age and gender classification of written text has been a popular field of research in natural language processing[1, 2, 3, 5]. The increasing popularity and accessibility of public blogs have brought attention to it in this area of research. Since the blog genre imposes no restrictions on the choice of topics and reflects authors not separated by geographical or demographic boundaries[1], our methods, if successful, can be extended to domain independent age attribution of authors.

Efforts in this area have several practical applications including but not limited to segment based opinion monitoring, blogger profiling, linking bloggers, business intelligence, facilitating customer discovery for businesses and enforcement of age restriction on blogging and social networking sites.

Dataset

We will be using the *Blog Authorship Corpus [J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006)]. Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.] available at: http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

The corpus is structured as a collection of blog posts from 19,320 bloggers totalling 681,288 posts (~140 million words) with roughly 35 posts (7250 words) per person.
Details pertaining to the blogger's age, gender, industry involved in etc. are included for each user.
In the corpus, users are divided by age as follows :

Age	**No. of bloggers**	**Age**	**No. of bloggers**	**Age**	**No. of bloggers**
13	690	26	1340	39	152
14	1246	27	1205	40	145
15	1771	28-32	0	41	139
16	2152	33	464	42	127
17	2381	34	378	43	116

18-22	0	35	338	44	94
23	2026	36	288	45	103
24	1895	37	259	46	72
25	1620	38	171	47-above	148

[Schler et al.,2006] mentions that gaps in the data for specific age ranges were created to avoid border cases. Nonetheless, for our purpose, we proceed with 3 classes (as mentioned in the dataset link), cleaned and divided uniformly as required, as below :

- 8240 “10s” blogs (ages 13-17)
- 8086 "20s" blogs(ages 23-27)
- 2994 "30s" blogs (ages 33-47)

Proposed Algorithm

Features based on blog content (Unigram, Bigram, Bigram+) will be used individually as well as in conjunction with shallow syntactic features (POS tags + Unigram) and deep syntactic features (PCFG based production rules). A classifier (SVM and NB) will be used to classify the blog posts into age groups of the authors, based on the features mentioned above. We will be using 5-fold nested cross validation method to test the efficacy of our features, which are summarized below:

Feature Encoding
Words: Unigram, Bigram, Bigram+
Shallow Syntax: POS tags + Unigram
Deep Syntax: Different encoding of production rules based on Probabilistic Context Free Grammar (PCFG) parse trees + Unigram, as done in [Song et. al., 2012]

Packages to be used

- SVM and Naive Bayes Classifiers
- Berkeley parser, for PCFG parsing **https://code.google.com/p/berkeleyparser/**

Programming Language & Operating System

Languages: Python, bash
Operating System: Linux (Ubuntu 12.04)

Baselines and Previous works

Since we will be classifying the bloggers into three age categories uniformly, the random baseline that we would be comparing against is **33.33%.**

We will also be comparing against the performance of a similar age group classification task done by *[Schler et. al, 2006]* based on style (POS, function words) and Content (LIWC categories) which reports an accuracy of **76.1%.**

The best results have been obtained by *[Rosenthal et. al. 2011]* who report an accuracy of **81.57%** using age based classification based on three features - online behavior, lexical content and lexical-stylistic elements of blog posts. They describe their efforts in age prediction of bloggers as a way to discern differences in blogging styles of pre- and post-social media era bloggers.

[Goswami et. al, 2009] report an accuracy of **80.1 %** for age group classification using sentence length, non-dictionary features and usage of slang words. It is worth mentioning that *[Rosenthal et. al, 2011]* show some skepticism, citing "*Goswami et al. (2009) add to Schler et al.'s approach using the same data and have a 4% increase in accuracy. However, the paper is lacking details and it is entirely unclear how they were able to do this with fewer features than Schler et al.*"

References

[1] J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

[2] Rosenthal, Sara, and Kathleen McKeown. "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations." *Proc. of ACL-11* (2011): 763-772.

[3] Carolyn, Dong Nguyen Noah A. Smith, and P. Rosé. "Author Age Prediction from Text using Linear Regression."

[4] Feng, Song, Ritwik Banerjee, and Yejin Choi. "Syntactic Stylometry for Deception Detection."

[5] Goswami, Sumit, Sudeshna Sarkar, and Mayur Rustagi. "Stylometric analysis of bloggers' age and gender." *Third International AAAI Conference on Weblogs and Social Media*. 2009.