GENDER AND AGE CLASSIFICATION ON THE BASIS OF BLOGS

[1]SAURABH SINGH, [2]MEGHNA MISHRA [3]SARWAN RATHORE, [4]MAYANK BHARSHIV, [5]ASHA SATHE

E-mail: saur.gbaa@gmail.com, meghnamish@gmail.com, karnot.sarwan@gmail.com, mayank.bharshiv@gmail.com, apsathe@yahoo.com

Abstract- Text is still the most prevalent Internet media type. Examples of this include popular social networking applications such as Twitter, Craigslist, Facebook, etc. Other web applications such as e-mail, blog, chat rooms, etc. are also mostly text based. A question we address in this paper that deals with text based Internet forensics is the following: given a short text document, can we identify if the author is a man or a woman? This question is motivated by recent events where people faked their gender on the Internet. Note that this is different from the authorship attribution problem. Analysis of a corpus of tens of thousands of blogs indicates significant differences in writing style and content between male and female bloggers as well as among authors of different ages. Such differences can be exploited to determine an unknown author's age and gender on the basis of a blog's vocabulary.

Keywords- Blogs, Support Vector Machine, corpus, information retrieval

I. INTRODUCTION:

The increasing popularity of publicly accessible blogs offers an unprecedented opportunity to harvest information from texts authored by hundreds of thousands of different authors. Conveniently, many of these blogs include formatted demographic information provided by the authors. (While much of this information is no doubt spurious, it is reasonable to assume that most is not.) Moreover, the blog genre imposes no restrictions on choice of topic. Thus, we can exploit this rich data set to begin to answer the following question: How do content and writing style vary between male and female bloggers and among bloggers of different ages? How much information can we learn about somebody simply by reading a text that they have authored? These are very basic questions that are both of fundamental theoretical interest and of great practical consequence in forensic and commercial domains. Accurate profiling of the demographics, background, and personality of an unknown author is a task of growing importance for national security, criminal investigations, and market research. We have shown how the right combination of linguistic features and machine learning methods enables an automated system to effectively determine several such aspects of an anonymous author; it is likely that other important profile components (such as educational background or other personality components) can also be extracted using such techniques, given appropriate training material. An important open research question, however, is the extent to which variation in genre and language might affect the nature of the models that can be used to solve various aspects of the profiling problem.

II. LITERATURE SURVEY

We investigate author gender identification for short length, multi-genre, content-free text, such as the ones found in many Internet applications. Fundamental questions we ask are: do men and women inherently use different classes of language styles? If this is true, what are good linguistic features that indicate gender? Experiments also indicate that function words, word-based features and structural features are significant gender discriminators. Psychology research suggests that one's state of mind, such as physical/mental health and emotion, can be gauged by the words he/she uses suggest that each author has a unique stylistic tendency, and refer to this feature as the author profile. Using these textual traces, researchers have begun to use online stylometric analysis techniques (authorship identification) as a forensic identification tool. But note that an author could conceals his/her true name by changing it to a name from the opposite sex. researchers distinguish between sex and gender, since sex is biologically founded, whereas gender is socially constructed (Crawford, 1995). Hence, by birth one's sex may be female or or male, but they are taught by society how to fulfill their corresponding gender. It needs to be made clear that all men are not masculine and all women are not feminine. Our research is based on the concept of gender-related language, rather than sex-related. For consistency purpose, we still use male/female to denote masculine/feminie. While a substantial literature has been devoted to isolating distinguishing characteristics of male/female linguistic styles (Lakoff 1975, Holmes 1993), this paper goes beyond earlier work in two ways: First, most of the previous work considered spoken language (Key 1972, Trudgill 1972, Labov 1990, Eckert 1997), which, unlike the formal written texts (i.e. books and articles) we consider, includes intonational, phonological and conversational cues. The relatively few studies on gender differences in writing have focused on more informal contexts – such as student essays (Mulac et al 1990, Mulac & Lundell 1994),

electronic communications (Herring 1996) and correspondence (Biber et al 1998, Palander-Collin 1999) – and some authors (Berryman-Fink & Wilcox 1983, Simkins-Bullock and Wildman 1991) have asserted that no difference between male and female writing styles in more formal contexts should be expected. There has been scant evidence thus far that differences between male and female writing are pronounced enough that they could be parlayed into an algorithm for categorizing an unseen text as being authored by a male or by a female.

Many different types of features have been considered as possible markers of textual style including lexical, syntactic, and vocabulary complexity-based features. For special cases, other feature types may e considered, such as grammatical or orthographic errors in unedited text, or morphological features or languages with especially rich morphological structure. However, the preponderance of evidence suggests that the most consistently effective features over a wide variety of authorship problems are function words and individual parts-of-speech. With the development of computers, stylometry has been widely accepted and has become dominant in identifying authorship. Over 1000 stylometric features have been proposed so far, including word- or character-based stylometric features (e.g. Yule, 1944; Holmes, 1992), function words (e.g. Mosteller and Wallace, 1984; Burrows, 1987), punctuation (e.g. Baayen et al., 2002), etc.

Although authorship identification methods have achieved some degree of success in many literary and forensic applications as mentioned above, very limited studies have been undertaken specifically for online messages.

There are also several other approaches to authorship identification. Statistical methods were widely used in earlier studies based on histograms of word-length distribution of various authors (Mendenhall, 1887), the Bayesian classifier (Mosteller and Wallace, 1984), principle component analysis (Burrows, 1987), cluster analysis (Holmes,1992), etc.

The advent of powerful computers instigated the extensive use of machine learning techniques in authorship analysis, such as decision tree (Apte et al., 1998), neural networks (Tweedie et al., 1996), support vector machine (SVM) (Diederich et al., 2000), etc.

III. APPROACH

To achieve the desired goal of the project the following approach is used:-

1. Text Categorization: In the text categorization problem, we are given a set of two or more categories and examples of texts in each category and are asked to correctly categorize unseen texts. the components of a text categorization system are now well understood:

   1. Document Representation – Choose a large set of text features which might potentially be
      useful for categorizing a given text (typically words that are neither too common nor too rare) and represent each text as a vector in which entries represent (some non-decreasing function of) the frequency of each feature in the text.
   2. Dimension Reduction – Optionally, use various criteria for reducing the dimension of the actors – typically by eliminating features which don't seem to be correlated with any category, by using latent semantic indexing , or by stepwise iteration of a learning algorithm .
   3. Learning Method – Use some machine learning method to construct one or more models of each category. Some of the most promising algorithms, which include k-nearest-neighbor, neural nets, Winnow, SVM can be used. (If multiple models are learned, methods such as bagging and boosting can be used to combine the models.)
   4. Testing Protocol – Finally, use some testing protocol, such as bootstrapping or k-fold cross validation to estimate the reliability of the system.

2. Stylometry:

While categorization by topic is typically based on keywords which reflect a document's content, categorization by author style uses precisely those features which are independent of content. Thus, stylometric models for categorization have typically been based on hand-selected sets of content-independent, lexical syntactic, or complexity-based features. Researchers in text categorization by topic typically use much larger feature sets, often in conjunction with automated feature selection methods .Some work in the stylometric community has also considered automated methods for selecting features.

3. Gender:

The object of this paper is to explore the possibility of automatically classifying formal written texts according to author gender. This problem differs from the typical text categorization problem which focuses on categorization according to topic. It also differs from the typical stylometric problem which focuses on authorship attribution – individual authors are more likely to exhibit consistent habits of style than large classes of authors. We will see, though, that using ideas from both the stylometric community and the text categorization community,

IV. SYSTEM FEATURES:

The product perspective is shown with the help of a work flow diagram.

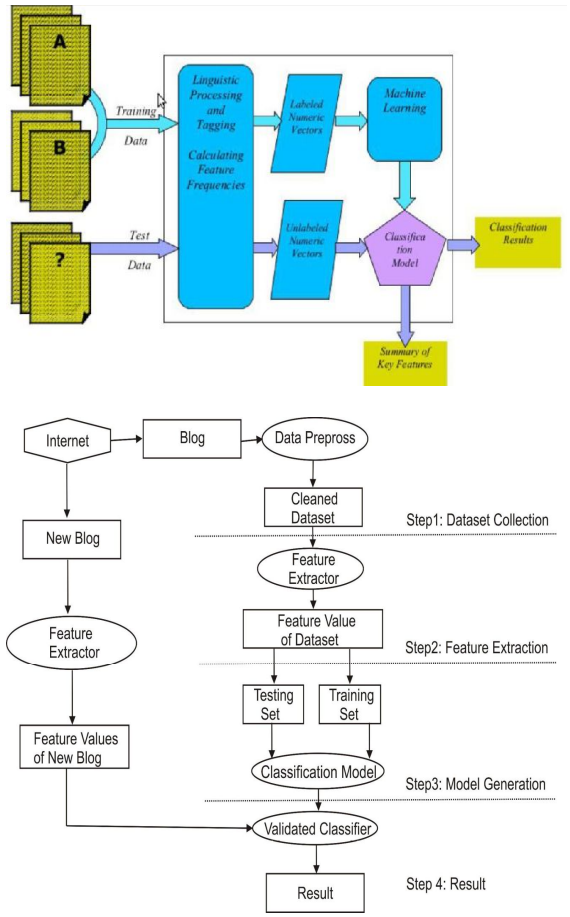


Figure 1 Gender identification process.

a) DATA PRE-PROCESSING:
Among various types of Internet text messages, newsgroup messages make use of neutral descriptive language while more private personal e-mails better reflect the true character of the author. Therefore, we use these two extreme types of datasets in the classifier design.

We retrieved the information of the authors, categorized the documents by the gender of the journalists, and discarded the documents written by authors whose gender could not be ascertained (e.g., neutral names). The text messages were then extracted by removing unnecessary information (such as date and time) and all XML formatting. Then we kept only the messages that contained more than 200 but less than 1000 words. Since most messages in the Reuters Corpus were newswire stories, these messages contained several quotes from others, which could bias the accuracy of the gender classifier. Therefore, we removed those messages that had too many quotations (") by setting the threshold of number of quotes/character counts to be 0.002.Table summarizes the corpus used in our analysis.

b)FEATURE SELECTION:
What are good linguistic features that indicate gender? This is an open research problem. Based on human psychology research (as discussed previously) and extensive experimen- tation, we classified five sets of gender-related features: (1) character-based; (2)word-based; (3) syntactic; (4) structure-based; and (5) function words.

c)AUTOMATIC FEATURE EXTRACTION AND REPRESENATION
For each message, the feature extractor produced a 545- dimension vector to represent the values of the 545 features. Since feature sets include information measured by various methods, the feature values we computed could range from 0 to more than 1000. For example, the first five feature values extracted from a message written by a male author could be 1420 0.988 0.0655 0.0119 0.217, which represents the total number of characters (C), the ratio of letters (total number of letters divided by C), the ratio of upper case characters (total number of upper characters divided by C), the ratio of digital characters (total number of digital characters divided by C) and the ratio of white spaces (total number of white-space characters divided by C), respectively.

To ensure all features are treated equally in the classifi- cation process, we normalized the features using max-min normalization method to ensure all feature values are between 0 and 1.

V. EXPERIMENTAL RESULTS

We conducted several experiments to examine the performance of different classification techniques, the impact of different combinations of parameters, and the significance of the proposed feature set. Feature extraction was implemented in Python and the classifiers were implemented in MATLAB Each experiment was conducted 10 times by 10-fold cross validation. Through extensive experiments we fixed Gamma ¼ 0.5 and C ¼ 3. We defined accuracy as the measure of prediction.

d) CLASSIFICATION TECHNIQUES:
For classifying we use either Bayesian-based logic regression, decision tree or a support vector machine. Thus, first, we applied the Bayesian-based logistic regression, Ada- Boost decision tree and SVM classifiers separately, using Reuters and Enron corpora. The results shown in indicate that SVM outperforms the other two methods for both datasets. The figure also reveals that the performance of the Bayesian logistic regression does not change significantly with the size of the training examples compared to the sharp improvement in AdaBoost (200 iterations) decision tree. The best classification result produced by the same classifier (SVM) was the accuracy of 76.75% and 82.23% respectively for two

different datasets, which implies the fact that the genders of the authors of neutral news (Reuters Newsgroup) are more difficult to discriminate than the personal e-mails (Enron Corpus).

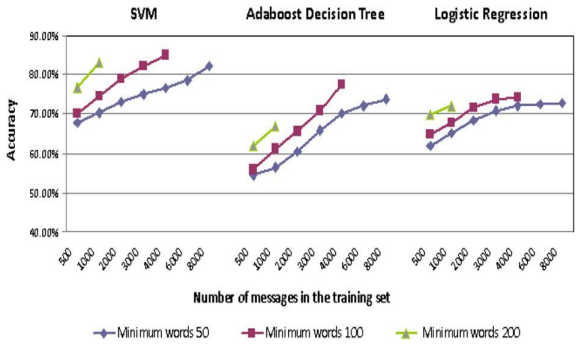


Figure 2 Comparative study of classification techniques

VI. MATHEMATICAL MODEL:

PROBLEM FORMULATION:

i)Author gender identification problem can be treated as A BINARY CLASSIFICATION PROBLEM i.e. given two classes {male,female}.assign a value depending upon whether the author is a male or female.

ii)Given two classes {male,female}assign an anonymous text message e, to one of these classes:

$e \in \{$ class 1,if author e is a male. / class 2,if author e is a female

iii)TRAINING SET: Design a set of fearless that remain relatively constant for a large no. of messages written by authors of the same gender

iv)A message is then represents by a d-dimensional vector where d is the total number of features.

v)This is then compared with the training data set and the category is determined.

AUTOMATIC FEATURE EXTRACTION AND REPRESENTATION (DATA PREPROCESSING):

To ensure that all features are treated equally in the classification process, we normalize the feature values are between zero and 1.

$$\text{Normalized } X_{ij} = \frac{X_{ij} - \min(X_{ij})}{\max(X_{ij}) - \min(X_{ij})}$$

X_{ij}=j^{th}feature in the i^{th}example.

$\max(X_{ij})$=max value of j^{th} feature.

$\min(X_{ij})$=min value of j^{th} feature.

→LEARNING ALGORITHM

Mathematically, we are interested in learning classifies y=f(x) from a set of TRAINING EXAMPLES to Create a set of TRAINING DATA:

Training examples=>

$$D=\{(x_1, y_1)(x_2, y_2) \dots\dots (x_N, y_N)\}$$

x_i = d – dimensional vector

$x_i = [x_{i1}, x_{i2}, \dots\dots x_{id}]$

$Y=\{y_i; i = 1,2,3, \dots\dots, N\}$denotes the label set, where

$Y_i \in \{+1, -1\}$

Class 1(-1)

Class 2(+1)

N:No. of compass in data set

→FEATURE SET SELECTION

We seek weight vectors W such that for each training document X, the vector dot product W*X excludes a threshold, if and only

We initially define two component weight vectors

$W^+=\{1,1,1,\dots..1\}$

& $W^-=\{-1,-1,\dots.-1\}$

Next , calibration of $W^+$and W^- using the following iterative procedure For each training example x, we have

c(x)=1 if x is female authored

c(x)=0 if x is male authored

i.e. S(W,X)=1 if W*X >0

S(W,X)=0 if W*X<0

W: weight vector

W_i:i^{th} element in W

X_i:i^{th} element in X

Then take examples one at a time and iteratively update the weight after each example, using formulae:

$$W_i^+ \leftarrow W_i^+ (1 + \beta x_i)^{(c(x) - S(W,X))}$$

$$W_i^- \leftarrow W_i^- (1 + \beta x_i)^{(c(x) - S(W,X))}$$

β : learning constant>0

∴Weights that improperly reduce the dot product are increased , and vice versa. Once all the examples have been used for training, they are randomly recorded and cycle of updates is run. This continues till all the training examples are correctly classified. Along the way, any elements of W^+ or W^- that drops below some threshold as set to zero. ∴Weights of features that appears most prominently in this classified documents are changed most dramatically. Thus, owing to this multiplicative update rule, weights of irrelevant features reduce to zero. ∴Learning algorithm is used for feature reduction: By iteratively running the learner,eliminating low weight features and rerunning, we can produce models with fewer and fewer features.

VII. CLASSIFICATION: SVM

Our objective here remains to find a linear optimal separating hyperplane. We find the optimal weight

vector w* by serving the following optimization problem

Minimize $j(w) = \frac{1}{2}||w||^2$

Subject to $y_i(w.x_i) \geq 1$

For the linear non separable cases, SVM bounds a soft margin by introducing the slack variable 2.

∴If the penalty function is linear,the problem now transforms to

Minimize $j'(w) = \frac{1}{2}||w^2|| + C\sum_{i=1}^{N} \in_i$

Subject to $y_i(w^* - x_i - b) \leq 1 - \in_1, \in_i \geq 0$

CONCLUSIONS:

We recognize that the problem of gender identification from text is an interplay between psycho-linguistics, generic writing styles of men and women, etc. Experimental results show that SVM outperforms the Bayesian-based logistic regression andAdaBoost decision tree for identifying author's gender from a given text document. By designing appropriate psycholinguistic and gender-linked features, we observe that wordbased features, function words an structural features play important roles in gender identification.

Experimental results indicate that the identification performance is improved by increasing the number of text documents in the training dataset as well as the number of words in each document (e-mail).Wefind that there are significant differences between men and women in personal writings such as e-mails, and gender differences also exist between authors of news articles even though neutral language is dominant there.

REFERENCES

[1] Argamon, S., M. Koppel , J. Fine and A. Shimoni (2003), Gender, Genre, and Writing Style in Formal Written Texts, Text , 23(3), 2003

[2] Baayen, H., H. van Halteren, F. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

[3] Biber, D. (1995). Dimensions of RegisterVariation: A Cross-linguistic Comparison (Cambridge University Press, Cambridge)

[4] de Vel O., M. Corney, A. Anderson and G.Mohay (2002). "Language and Gender Author Cohort Analysis of E-mail for Computer Forensics", Digital Forensic Research Workshop, August 7 – 9, 2002, Syracuse, NY

[5] Koppel, M., S. Argamon, A. Shimony (2002). Automatically categorizing written texts by author gender, Literary and Linguistic Computing 17,4, (2002), 401-412.

[6] Mosteller, F. and Wallace, D. L. (1964). Inference and Disputed Authorship: The Federalist. Reading, Mass. : Addison Wesley, 1964.

[7] Mulac, A., J.J. Bradac and P. Gibbons (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences, Human Communication Research 7:121-152 (2001)

[8] Pennebaker, J.W., & Stone, L.D. (2003). Words of wisdom: Language use over the lifespan. Journal of Personality and Social Psychology, 85, 291-301 (2003).

[9] Pennebaker, J.W., Francis, M.E., and Booth, R.J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Erlbaum Publishers

[10] Stamatatos, E., N. Fakotakis & G. Kokkinakis (2001).Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, (2001) 193—214.

[11] Yule, G.U. (1938). On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship, Biometrika, 30, (1938) 363-390.

★★★