Authorship Profiling in Gender Identification on English editorial documents using Machine Learning Algorithms

Raju. Nadimpalli. V.G.,
Associate Professor, Dept. of CSE, GRIET, Hyderabad, India

Gopala Krishna. P.
Associate Professor, Dept. of IT, GRIET, Hyderabad. Telangana.

Yelleni Mounica
Dept. of IT, GRIET, Hyderabad. Telangana.

V Sahithi
Dept. of IT, GRIET, Hyderabad, Telangana

Abstract— *Authorship analysis deals with the classification/identification of texts into classes based on the stylistic choices of their authors. Author profiling distinguishes between classes of authors studying their sociolect aspects. This helps in identifying profiling aspects such as gender, age, native language, or personality type. The present paper identifies the gender from the stylistic choices/features made by the people on English editorial documents by using supervised machine learning algorithms. The present paper achieves average performance of 93 to 944% using Random Forest classifier in identifier in identifying gender of an unknown editorial document.*

Keywords— authorship analysis; authorship profiling; supervised machine learning; forensic investigation;

I. INTRODUCTION

The rapid growth of the electronic documents in internet in the form of emails, blogs, social networking, news groups, twitter, Facebook, etc. has created multitude ways to share information across the World Wide Web. The main reason for this is rapid development and proliferation of internet technologies at very low cost. The digital audience engaged with newspaper content reached a new peak in January 2016, totaling 183 million adult unique visitors [1]. This phenomenal growth of accessing information has created problems in author identification profiling, because some people circulate some of the articles and sometimes combine two or more articles in the social media. Hence authorship profiling has become an emerging research area in Information retrieval research [2].

Author profiling distinguishes between classes of authors by studying their sociolect aspect, i.e., how language is shared or how an author can be characterized from a psychological viewpoint. This information helps in identifying profiling aspects such as gender, age, native language, or personality type [2]. Author profiling is a problem of growing importance, among others for applications in forensics, security, and marketing. From a forensic linguistics perspective, for example, one would like to learn about the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). From a marketing viewpoint, companies may be interested to learn about the demographics of people who like or dislike their products, given blogs and online product reviews as analysis source.

II. LITERATURE REVIEW

Many earlier researchers worked on authorship identification and authorship profiling using various linguistic features. The linguistic features are represented in terms of lexical, syntactic and semantic levels. The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as psychology, linguistics and, more recently, computational linguistics. Pennebaker [3] investigated how the style of writing is associated with personal attributes such as age, gender and personality traits, among others. Argamon et al. [4] investigated the task of gender identification on the British National Corpus and achieved approximately 80% accuracy. Similarly in [5] the authors investigated age and gender identification on formal texts. On the other hand, Zhang et al. [6] experimented with short segments of blog post and obtained 72.1% accuracy for gender prediction. Similarly, Nguyen et al. [7] studied the use of language and age among Dutch Twitter users. Since 2013, a lot of relevant research has been published in the context of the shared task on author profiling organized at PAN [8, 9, 10, 11, 12]. Koppel et al. [10, 15] studied the problem of automatically determining an author's gender by proposing combinations of simple lexical and syntactic features, and achieving approximately 80% accuracy. Schler et al. [13] studied the effect of age and gender in the writing style in blogs; they gathered over 71,000 blogs and obtained a set of stylistic features like non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain. Goswami et al. [14] added some new features as slang words and the average length of sentences, improving accuracy to 80.3% in age group identification and to 89.2% in gender detection.

III. METHODOLOGY

The detailed methodology is given below in terms of various steps.

Step 1: Document collection: - For gender classification purpose, a collection of 2000 editorial English documents are collected from the internet.

Step 2: Pre-processing: The step two performs preprocessing by making the corpus as case insensitive, and performs other operations like data cleaning, tokenizing, normalization, for effective feature extraction. By this step, spaces, numbers, special characters from the corpus articles are eliminated. We did not removed of stop words and stemming mechanisms.

Step 3: Feature Extraction: The features considered for the gender identification are lexical, character, function and gender specific feature words. Lexical features include word length, sentence length, word frequencies, vocabulary richness functions, word n-grams etc. Character features include frequency of character types, frequency of character n-grams. Most lexical features are highly author and language dependent. Totally 150 function words and 236 gender specific feature words are used for gender identification. They are *"a an the yes no okay OK all everybody his most other that what your another everyone I much others theirs whatever yours any everything it myself ours them which yourself anybody few its neither ourselves themselves whichever anyone he itself no one several these who both hers many none some this whom anything her little nobody she they whoever both hers many none some this whom each herself me nothing somebody those whomever each other him mine one someone us whose either himself more one another something we you are can didn't hadn't haven't might shouldn't won't aren't cannot do 'd 've mightn't was 'll ain't can't don't has is mustn't wasn't would 're could does hasn't isn't shall were wouldn't be couldn't doesn't 's 's shan't weren't 'd be couldn't doesn't 's 's shan't weren't 'd and or though now that if while in order that in case because yet unless even though now that whereas even if nor so when although only if whether or not until adios bah dear Ha-ha howdy oops tush whoosh ah begorra doh hail hoy ouch tut wow aha behold duh hallelujah huh phew Tutetut yay ahem bejesus eh heigh-ho humph phooey ugh yikes ahoy bingo encore hello hurray pipepip uh-huh yippee alack bleep eureka hem hush pooh uh-oh yo alas boo fie hey indeed pshaw uheuh yoicks all hail bravo gee hey presto jeepers creepers rats viva yoo-hoo alleluia bye gee whiz hi jeez righto voila yuk aloha cheerio gesundheit hip lo and behold scat wahoo yummy amen cheers goodness hmm man shoo well zap attaboy ciao gosh ho my word shoot whoa ay cripes hah hot dog ooh Touch whoops aw crikey great ho hum now so long whoopee aboard astride down of through worth about at during off throughout onto above athwart except on till absent atop failing across barring following opposite toward after before for out towards against behind from outside under along below in over underneath alongside beneath inside past unlike amid beside into per until amidst besides like plus up among between mid regarding upon into amongst beyond minus round via around but near save with as by next since within aslant despite notwithstanding than without aslant despite notwithstanding than without"*

Step 3: **Evaluates 250 Most Frequently Words (MFW)** highest mutual information using our java program. This allows to test the relevance of selecting MFW as clues for Authorship Attribution

Step4: **Vector Space Model Representation**: A two dimensional matrix is constructed, that exhibits the frequency occurrences of the 250 MFW from all the documents of the authors and assign class label.

Step5: **Gender prediction**: Using machine learning classifiers, gender of the unknown document is predicted. The classifiers used are logistic regression, support vector machines, naïve Bayes, decision trees, logitboost, and random forest using Weka 3.7 data mining software for Machine Learning.

IV. RESULTS AND DISCUSSION

The present paper is implemented on a collection of 2000 editorial documents from twenty English leading editorial columnists of India. Out of 20 columnists 10 are male rest are female columnists. They are (1) Alia Allana, (2) Chetan Bhagat, (3) Barkha Dutt(4) Shobhaa De (5) Sangeetha Devi Dundoo (6) Chandra Sekar (7) Josy Joseph (8) Suresh Menon(9) Renuka Bisht and (10) Sainath. The editorials are collected from the leading newspapers of India namely The Hindu, Times of India and Sunday Guardian. Nearly 100 Documents per author has been considered for both training and testing purpose.

A java module has been implemented for extraction of lexical, character, function and feature words from the editorial documents, calculating 250 most frequent features, representation of vector space model representation of frequent features and editorial documents. Then using Weka 3.7 for the sake of supervised machine learning algorithms. The algorithms considered in this paper are Logistic Regression, Support Vector Machines, LogitBoost, Naïve Bayes, Decision Trees, Random Forest classifiers. Table 1

shows the results of gender identification of various algorithms on both female and male editorial column data.

Table 1: Gender identification on supervised machine learning algorithms

Male/Female	Logistic	SMO	Logit Boost	Navie Bayes	Decision Trees	Random Forest
Male (99,1)	100	100	100	100	100	100
Male(95,5)	80	100	100	100	100	100
Male(90,10)	80	80	60	60	90	90
Male(50,50)	86	88	80	80	80	82
Female(99,1)	100	100	100	100	100	100
Female(95,5)	100	100	100	100	100	100
Female(90,10)	60	100	100	100	70	100
Female(50,50)	76	86	76	84	78	90

The experiment is conducted on Weka 2.7 tool for predicting gender of an unknown documents on various supervised machine learning classifiers on both training and testing data. The classifiers used for gender prediction are: LogitBoost, class performs classification using a regression scheme as the base learner, and can handle multi-class problems. Sequential Minimal Optimization (SMO) by John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data. Logistic Regression**,** class for building and using a multinomial logistic regression model with a ridge estimator. NaiveBaye**s, c**lass for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier. Decision Trees, Class for generating a pruned or unpruned C4.5 decision tree. RandomForest, Class for constructing a forest of random trees.

The data has been divided into training and testing. The table1 shows Male (99, 1) means 99 training documents and 1 test document has been taken for the experimentation. The experiment is carried out on a 10 fold cross validation, where 10 fold cross validation refers to break data into 10 sets of size n/10, then train on 9 datasets and test on 1 and repeat 10 times and take a mean accuracy.

From the table 1 Random Forest classifier outperforms all other algorithms with an average of 93% for male authors and 94% for female authors. Then SVM SMO classifier performed well with 92% for male and 938% for female authors. The figure 1 shows average gender identification of various algorithms on Male and Female authors. From the figure 1 it is evident that the identification of female author's percentage is higher than male authors.

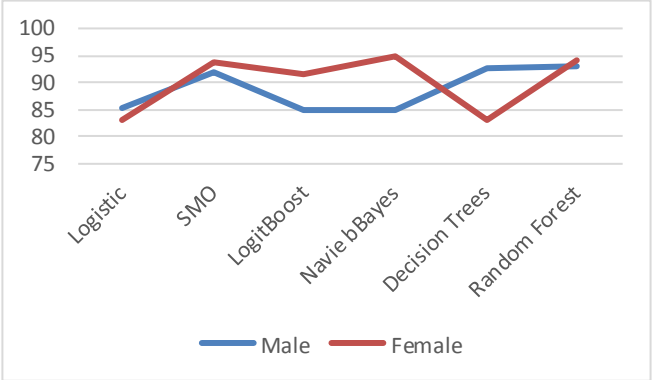

Fig. 1: Average % of gender identification on Male and Female data

V. CONCLUSIONS

Author profiling is used to identify the gender, age, native language, or personality type. The present paper identifies the gender from the stylistic choices/features made by the people on English editorial documents using supervised algorithms. The present paper achieves average performance of 93 to 944% using Random Forest classifier in identifier in identifying gender of an unknown editorial document. Author profiling is a problem of growing importance in applications in forensics, security, and marketing.

VI. REFERENCES

1. Vijay Kumar, Ganapathi Raju, "Histograms of Term Weight Feature (HTWF) model for Authorship Attribution", Volume 10, Number 16 (2015) pp 36622-36628, IJAER, 2015.
2. http://pan.webis.de/clef17/pan17-web/author-profiling.html
3. James W. Pennebaker, Mathias R. Mehl, and Kate G. Niederhoffer, "Psychological aspects of natural language use: our words, our selves". Annual review of psychology, 54(1):547–577, 2003.
4. Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni.Gender, genre, and writing style in formal written texts. TEXT, 23:321–346, 2003.
5. John D. Burger, John Henderson, George Kim, and Guido Zarrella.Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

6. Cheng, Na, et al. "Gender identification from e-mails." *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*. IEEE, 2009.
7. Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "how old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.
8. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180, 2014.
9. Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Pottast, Benno Stein. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Balog K., Capellato L., Ferro N., Macdonald C. (Eds.) CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1609, pp. 750-784
10. Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically categorizing written texts by author gender." *Literary and Linguistic Computing* 17.4 (2002): 401-412.
11. Murugaboopathy, G., et al. "Appropriate gender identification from the text." International Journal of Emerging Research in Management and Technolog (2013): 58-61.
12. Madhulika Agrawal and Teresa Gonçalves. Age and gender identification using stacking for classification.
13. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pages 199–205. AAAI, 2006.
14. Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. Stylometric analysis of bloggers' age and gender. In Eytan Adar, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, ICWSM. The AAAI Press, 2009.
15. Rangel, Francisco, et al. "Overview of the 2nd author profiling task at pan 2014." CEUR Workshop Proceedings. Vol. 1180. CEUR Workshop Proceedings, 2014.