Content-centric Age and Gender Profiling

Lim Wee Yong, Jonathan Goh, Vrizlynn Thing
Cybercrime & Security Intelligence Department,
Institute for Infocomm Research, Singapore

Introduction

Author profiling is a form of text analysis where the objective is to ascertain characteristics of the author behind a text sample. The aim of this task is to classify text samples into:

- Gender
- Age groups
- On both English and Spanish corpora

Challenges

- A person's syntactic construct or lexical usage can give cues to authorship, but what features should be used?
- No consensus on "ideal" features to use
- High dimensional and difficult to find "top" words

Contributions

- Proposal of new content based Feature
- Comparison of content-based features with some other common features widely used in this area
- Selection for the best features and the proper use of classification
- Measure similarity among content and profile group's samples

Prior Work

- Koppel et al. (2003, 2009), Juola (2006), Labbe (2007), Zheng et al. (2006) provided many authorship features

Types of Features

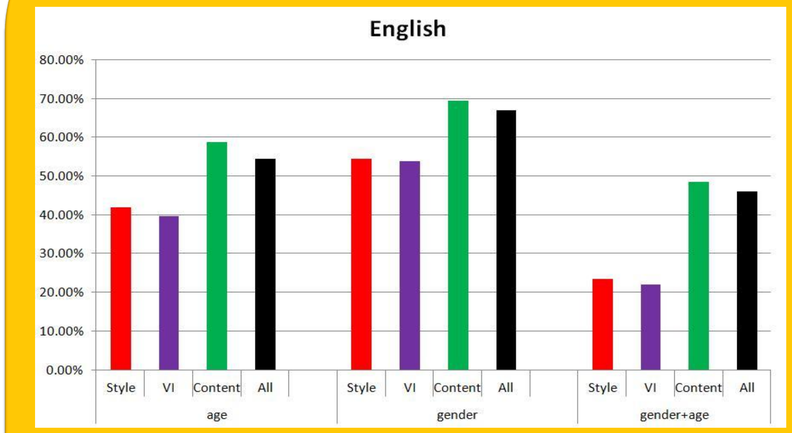
- Style-Based Features
  - pronouns, determiners & prepositions
  - average sentence length, words per conversations, number of contraction words & URLs
- Vocabulary and Idiosyncrasies Features
  - unique words

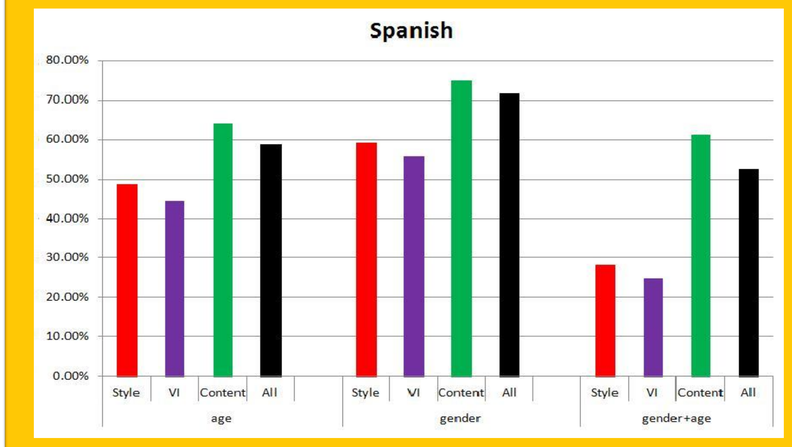
Our Novel Features

- Content-based features
  - Reflective of subject areas expressed in conversations
  - *n* most discriminative common words for each profile group

"Why do authors in different sociolinguistic profile group differs in their written communications, assuming using a common language?"

Experiment Results (Accuracy vs Features)





Conclusions

- Novel and concise (low dimensionality) content-based feature
- Superior performance of content-based features compared to some common features used in this area

References:
Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17, 401–412 (2003)
Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: In IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (2003)
Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3) (2006)
Labbé, D.: Experiments on authorship attribution by intertextual distance in English.Journal of Quantitative Linguistics 14 (2007)
Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3) (2006)