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

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: