Using Stylometric Features to Predict Author Personality Type in Modern Greek Essays

Gagiatsou Sofia, Markopoulos Georgios, Mikros George

Gagiatsou Sofia
Department of Linguistics, School of Philosophy
National and Kapodistrian University of Athens
Athens, Greece
e-mail: sgagiats@phil.uoa.gr
Gagiatsou Sofia holds a PhD in Linguistics (National and Kapodistrian University of Athens-UoA, 2021, Thesis: Automatic author profiling based on natural language processing techniques). She holds a MSc in Language Technology (National Technical University of Athens & UoA, 2006) and a BA in Greek Philology and Linguistics (UoA, 2002). She has been involved in various R&D projects. Her research interests concern language resources and the development and processing of textual corpora.
Introduction

- Authorship identification
  - Authorship Attribution
  - Authorship Verification
  - Authorship Profiling
    - Computational Personality Prediction (CPP)

➢ Our contribution:

We performed the first CPP study in Modern Greek focused on high-school students.
Development of the optimal classification model in order to predict author's personality based on natural language processing techniques applied to essays written in Modern Greek by high-school students.

Each writer has been profiled by filling in the Jung Typology Test.

The feature set employed was stylometric.

Machine learning algorithms were ranked according to their cross-validated accuracy.
Related Works

Carl Jung’s and Isabel Briggs Myers’ personality type theory.

Personality Research from Text.


B. Plank and D. Hovy, “Personality Traits on Twitter-or how to get 1,500 Personality Tests in a Week” The Sixth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2015, pp. 92-98.


Methodology

- Corpus development.
- Stylometric features extraction.
- Classification and prediction algorithms selection and application.
Corpus: a unique dataset.

- Collection of primary textual data from native speakers of Modern Greek.
- The corpus consists of essays of 198 high school students and comprises 250,000 words in total.
- Corpus pre-processing with natural language processing tools (tokenizer, lemmatizer, and POS tagger).
Methodology

Stylometric features extracted

- the most frequent character bigrams, and trigrams
- the most frequent words bigrams, and trigrams
- mean word and sentence length
- the occurrence frequency of content and functional words
- the most and less frequent words
- the occurrence frequency of parts of speech
- hapax and dis legomena
### Classification Algorithms

1. Naive Bayes  
2. Generalized Linear Model  
3. Logistic Regression  
4. Fast Large Margin  
5. Deep Learning  
6. Decision Tree  
7. Random Forest  
8. Gradient Boosted Trees  
9. Support Vector Machine
The best results were obtained by the Naive Bayes algorithm.

<table>
<thead>
<tr>
<th>Personality Type</th>
<th>Naive Bayes Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Extraversion</td>
<td>80.7%</td>
</tr>
<tr>
<td>Intuition</td>
<td>79.9%</td>
</tr>
<tr>
<td>Feeling</td>
<td>68.8%</td>
</tr>
<tr>
<td>Judging</td>
<td>75.7%</td>
</tr>
</tbody>
</table>
Weights for Extraversion.

Mean length of sentence, words that occur only twice in one text, the most frequent content words, and personal pronouns.

Weights for Intuition.

Word's mean length, the most frequent trigrams of characters, hapax legomena, personal pronouns, content words, the most frequent word bigrams, the rarest words, the most frequent word trigrams, and all content words.
Weights for Feeling.

Verbs, adjectives, the most frequent content words, personal and possessive pronouns, nouns, and adverbs.

Naive Bayes - Weights

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>VerbsFreq</td>
<td>0.270</td>
</tr>
<tr>
<td>AdjectivesFreq</td>
<td>0.216</td>
</tr>
<tr>
<td>PercentageOfTopMostFreqBigramsStopWordsCoveragFile</td>
<td>0.200</td>
</tr>
<tr>
<td>PersonalAndPossessivePronounsFreq</td>
<td>0.096</td>
</tr>
<tr>
<td>NounsFreq</td>
<td>0.084</td>
</tr>
<tr>
<td>AdverbsFreq</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Weights for Judging.

The most common word trigrams, the most common word bigrams, mean length of sentence, the most common character bigrams and the most common character trigrams, personal and possessive pronouns, articles, and word's mean length.

Naive Bayes - Weights

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>PercentageOfTopMostFreqTrigramsCoverageFile</td>
<td>0.763</td>
</tr>
<tr>
<td>PercentageOfTopMostFreqBigramsCoverageFile</td>
<td>0.529</td>
</tr>
<tr>
<td>AverageSentenceLength</td>
<td>0.776</td>
</tr>
<tr>
<td>PercentageOfTopMostFreqCharBigramsCoverageFile</td>
<td>0.101</td>
</tr>
<tr>
<td>PercentageOfTopMostFreqCharTrigramsCoverageFile</td>
<td>0.101</td>
</tr>
<tr>
<td>PersonalAndPossessivePronounsFreq</td>
<td>0.077</td>
</tr>
<tr>
<td>ArticlesFreq</td>
<td>0.081</td>
</tr>
<tr>
<td>AverageWordLength</td>
<td>0.065</td>
</tr>
</tbody>
</table>
The reported results show a competitive approach to the personality prediction problem.

The research revealed new combinations of stylometric features and corresponding computational techniques, giving interesting and satisfying solutions to the problem of the author’s personality prediction for Modern Greek.
Future work

- Experimentation with more linguistic features.

- Use of well-known psychometric lexicons in Modern Greek to further enrich our feature sets.
Thank you!