The Fifteenth International Conference on Digital Society
ICDS 2021
July 18, 2021 to July 22, 2021 - Nice, France

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

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

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:

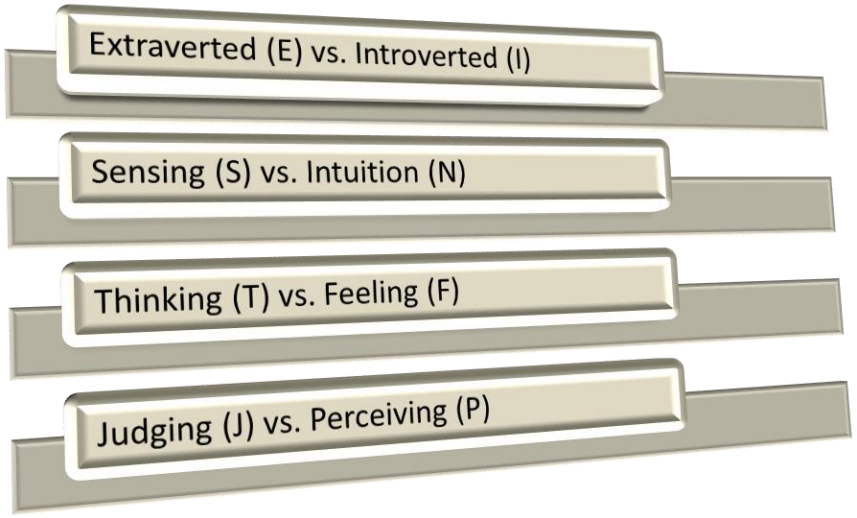
We performed the first CPP study in Modern Greek focused on high-school students.

Introduction

- 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.

K. Luyckx and W. Daelemans, "Personae: A corpus for Author and Personality Prediction from Text" The Sixth International Language Resources and Evaluation Conference (LREC 2008), 28-30 May 2008, pp. 2981-2987.

D. Brinks and H. White, "Detection of Myers - Briggs Type Indicator via Text based Computer-mediated Communication," CS 229 Machine Learning Projects, Stanford, 2012.

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.

B.Verhoeven, W. Daelemans, and B. Plank, "Creating TwiSty: Corpus Development and Statistics," Computational Linguistics and Psycholinguistics Research Center CLiPS Technical Report Series, University of Antwerp, Belgium, CTRS-006, 2016.

L. C. Lukito, A. Erwin, J. Purnama, and W. Danoekoesoemo, "Social Media User Personality Classification using Computational Linguistic" The Eighth International Conference on Information Technology and Electrical Engineering, Oct. 2016, pp. 1-6.

K. Yamada, R. Sasano, and K. Takeda, "Incorporating Textual Information on User Behavior for Personality Prediction" The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Jul.-Aug. 2019, pp. 177-182.

Methodology

- Corpus development.
- Stylometric features extraction.
- Classification and prediction algorithms selection and application.

Methodology

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).
- The output was submitted to the data mining platform Rapidminer [Mierswa, I., & Klinkenberg, R. (2018). RapidMiner Studio (9.1). Data science, machine learning, predictive analytics. Retrieved from https://rapidminer.com/.]

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

Methodology

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

Results

The best results were obtained by the Naive Bayes algorithm.

Personality Type	Naive Bayes Classifier		
	Accuracy	Precision	Recall
Extraversion	80.7%	80.5%	100%
Intuition	79.9%	81.3%	92.6%
Feeling	68.8%	67.7%	96.7%
Judging	75.7%	76.2%	95.2%

Results

Weights for Extraversion.

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

Naive Bayes - Weights

Attribute	Weight
ActiveVoiceVerbsFreq	0.243
AverageSentenceLengthInWords	0.163
PercentageOfTokensAppearingTwiceCoverageInFile	0.100
PercentageOfTopMostFreqNonStopWordsCoverageInFile	0.079
PersonalPronounsFreq	0.078

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.

Naive Bayes - Weights

Attribute	Weight
AverageWordLength	0.227
PercentageOfTopMostFreqCharTriGramsCoverageInFile	0.181
PercentageOfTokensAppearingOnceCoverageInFile	0.173
PersonalPronounsFreq	0.170
NumOfSingleNonStopWordsPerAllWordsOccurrencesInFile	0.156
PercentageOfTopMostFreqBiGramsCoverageInFile	0.116
PercentageOfBottomLeastFreqTokensCoverageInFile	0.100
PercentageOfTopMostFreqTriGramsCoverageInFile	0.080
PercentageOfAllNonStopWordsCoverageInFile	0.017

Results

Weights for Feeling.

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

Naive Bayes - Weights

Attribute	Weight
VerbsFreq	0.270
AdjectivesFreq	0.219
PercentageOfTopMostFreqNonStopWordsCoverageInFile	0.200
PersonalAndPossessivePronounsFreq	0.096
NounsFreq	0.084
AdverbsFreq	0.048

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

Attribute	Weight
PercentageOfTopMostFreqTriGramsCoverageInFile	0.352
PercentageOfTopMostFreqBiGramsCoverageInFile	0.329
AverageSentenceLengthInWords	0.226
PercentageOfTopMostFreqCharBiGramsCoverageInFile	0.101
PercentageOfTopMostFreqCharTriGramsCoverageInFile	0.101
PersonalAndPossessivePronounsFreq	0.077
ArticlesFreq	0.061
AverageWordLength	0.061

Conclusions

- 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!