EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Graduate Program in Information Systems
PPgSI

6th PPgSI's Dissertations Workshop 2019

Authorship attribution with authorship characterization techniques

Author:	Caio Deutsch			
Advisor:	Dr. Ivandré Paraboni			
Co-advisor:	NA			
Research lines:	[ ] Systems Development and Management		[X] Systems Intelligence	
Research areas:	[ ] Database [ ] Information technology management	[ ] Software engineering [ ] Human-Computer Interaction	[X] Artificial intelligence [X] Pattern recognition	[ ] Graphics processing [ ] Optimization
Application areas:	[ ] Enterprise environments / Business processes [ ] Economy [ ] Games / Serious games [ ] Health	[ ] Bioinformatics [ ] Education / Distance learning [X] Linguistics / Natural Language [ ] Other Which? _____	[ ] Biometrics [ ] E-government [ ] Cheminformatics	[ ] Mobile devices [X] Internet / Social Networks [ ] Robotics [ ] General*
Period in the program (at the workshop date):	[ ] 2nd semester	[X] 3rd semester	[ ] 4th semester	[ ] 5th semester
Qualifying:	[ ] Qualifying held in: dd/mm/yyyy		[X] Plan for qualifying in: 15/09/2019	
Defense:	Deadline for deposit: dd/mm/yyyy		Plan for defending in: 15/06/2020	
Publications associated with the master's project:	**No publications to date.**			

The research project summary

Context:

Authorship attribution (AA), a sub-area of Natural Language Processing (NLP), seeks to identify the author of a given text in a set of possible authors.

This task of identifying authors has been studied by experts in the field for at least six decades, where the analyzed texts were still handwritten and not digital as in the present day.

Technological advancement has led to the creation of large data sources for NLP studies, specifically for the AA problem. These studies have been addressing areas such as news, where so-called fake news cause controversy on common themes in our current society, such as politics and state elections.

Research problem:

Author attribution is increasingly showing importance for various social activities, especially forence analysis. Applications can help, for example, unravel mysteries of fake news authors, authoring source code, or even identifying aliases.

Studies involving AA demonstrate modest results and motivate the exploration of different techniques to improve the accuracy of current models. This is the case of (UCELAY et al., 2016), where the authors used authoritative characterization (CA) techniques to predict gender and age. These models could be used to solve the AA problem.

In addition, the problem of AA can be approached for several languages, however the vast majority of studies are focused on the English language, and studies in Brazilian Portuguese are rare.

Research objective:

Develop models capable of identifying authors of a particular text using characterization authorship (CA) techniques, in order to obtain superior results to traditional AA models.

Characteristics of the proposed solution:

The classification model is composed of a set of three other classifiers: n-grams of characters patterns, n-grams of non-diacritically distorted characters, and n-gram words and a binary variable which simulates a four classifier which will be a CA classifier in the next experiment. The four variables are ensemble with a logistic regression and used to predict a AA problem.

Theoretical foundations:

- Part of speech
- Text distortion
- Word embeddings

Correlated works:

- CUSTODIO, J. E.; PARABONI, I. Each-usp ensemble cross-domain authorship attribution: Notebook for pan at clef 2018
- UCELAY, M.; VILLEGAS, M.; FUNEZ, D.; CAGNINA, L.; ERRECALDE, M.; RAMIREZ-DE-LA-ROSA, G.; VILLATORO-TELLO, E. Profile-based approach for age and gender identification
- ROCHA, A.; SCHEIRER, W. J.; FORSTALL, C. W.; CAVALCANTE, T.; THEOPHILO, A.; SHEN, B.; CARVALHO, A. R. B.; STAMATATOS, E. Authorship attribution for social media forensics.

Validation

The experiments will be validated using conventional machine learning measures such as recall, accuracy, precision, and F measurement.

Limitations, risks, and threats:

The scope of this project is limited to AA problems using CA techniques with machine learning methods. Computational methods based on complex networks, graphs and compression models will not be considered. The mentioned corpus are limited to domains of literature, Social networks such as, facebook, twitter, whatsapp and others. The languages considered are Brazilian Portuguese and English.

Scientific contribution:

This research aims to advance the frontier of knowledge about the problem of AA using CA techniques using language and domain independent models. By developing and studying AA models using CA techniques, it is expected to advance knowledge of the relationship between language and the factors that determine authorship in both areas. In particular, it is expected to advance studies to the Portuguese language.

The research method			
Genre (choose ONE)	[ ] Theoretical research	[X] Practical research	[ ] Empirical research [ ] Methodological research
Nature (choose ONE)	[ ] Basic research	[X] Applied research	
Approach (choose ONE)	[ ] Quantitative research	[ ] Qualitative research	[X] Quali-quanti research
Literature review* (you can choose more than one)	[ ] Narrative review [X] Descriptive review [ ] Scoping review	[ ] Meta-analysis [ ] Qualitative systematic review [ ] Umbrella review	[ ] Theoretical review [ ] Realistic review [ ] Critical review
Main technical procedure (choose ONE)	[ ] Experimental research [ ] Bibliographic research [ ] Documental research [ ] *Ex-post-facto* research	[ ] *Survey* [X] Case study [ ] Participatory research [ ] Research-action	[ ] Ethnographic research [ ] Grounded theory [ ] Design science [ ] Other Which? _____
Data analysis (you can choose more than one)	[X] Descriptive statistics [X] Inferential statistics	[ ] Statistical test [X] Content analysis	[ ] Discourse analysis [ ] Others: _____

* Definition of types of literature reviews established by Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Next steps:

- Finish the text for the qualification exam
- Develop a CA classifier to predict de binary variable used in experiment 1
- Write down the results of this new experiment
- Schedule the defense