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

Programa de Pós-graduação em Sistemas de Informação
PPgSI

VI Workshop de Dissertações de Mestrado do PPgSI
2019

Caracterização autoral a partir de textos utilizando redes neurais artificiais

Autoria de:	Rafael Felipe Sandroni Dias			
Orientação de:	Profa. Dr. Ivandré Paraboni			
Coorientação de:				
Linha de pesquisa:	[ ] Gestão e Desenvolvimento de Sistemas		[x] Inteligência de Sistemas	
Área de pesquisa:	[ ] Banco de dados [ ] Engenharia de software [ ] Gestão de tecnologia da informação [ ] Interação humano computador		[x] Inteligência artificial [ ] Processamento gráfico [ ] Reconhecimento de padrões [ ] Otimização	
Área de aplicação:	[ ] Ambientes corporativos / Processos de negócio [ ] Bioinformática [ ] Biometria [ ] Dispositivos móveis [ ] Economia [ ] Educação / Educação a distância [ ] Governo eletrônico [ ] Internet / Redes sociais [ ] Jogos / Jogos sérios [x] Linguística / Língua natural [ ] Quimioinformática [ ] Robótica [ ] Saúde [ ] Outra Qual? _____ [ ] Geral*			
Semestre no curso (na data do workshop):	[ ] 2º semestre	[ ] 3º semestre	[ ] 4º semestre	[x] 5º semestre
Qualificação:	[x] Qualificação já realizada em: 29/06/2018		[ ] Realização da qualificação planejada para:	
Defesa:	Prazo máximo para depósito: 03/09/2019		Realização da defesa planejada para: 03/09/2019	
Publicações associadas ao projeto de mestrado:	• Publicado: Rafael Felipe Sandroni Dias; Ivandré Paraboni (2018) Author profiling using word embeddings with subword information. 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter (PAN-CLEF 2018) • Publicado: Rafael Felipe Sandroni Dias; Ivandré Paraboni (2018) Caracterização autoral de usuários do facebook: gênero, idade, religiosidade e área de formação. I Congresso Internacional em Humanidades Digitais (HDRio-2018). • Publicado: Hsieh, Fernando Chiu; Rafael Felipe Sandroni Dias; Ivandré Paraboni (2018) Author Profiling from Facebook Corpora. 11th International Conference on Language Resources and Evaluation (LREC-2018) pp. 2566-2570. Miyazaki, Japan.			

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

Programa de Pós-graduação em Sistemas de Informação
PPgSI

Resumo do projeto de pesquisa

Contexto:

A caracterização autoral (do inglês, Author Profiling) é uma tarefa computacional de reconhecimento de características sociais (i.e., caracterização de gênero, faixa etária, grau de educação etc) de autores a partir de textos. Na literatura, modelos de caracterização autoral utilizam, em sua grande maioria, conhecimentos linguísticos especializados para cada tipo de tarefa, idioma e domínio de texto. Contudo, estudos recentes de processamento de língua natural (PLN) têm obtido resultados satisfatórios com uso de redes neurais artificiais para extração de informações sintáticas e semântica em problemas de classificação de textos.

Problema de pesquisa:

Os modelos computacionais de caracterização autoral tendem a utilizar diversos métodos baseados em conhecimento linguístico prévio. No entanto, em diversas tarefas de PLN, os métodos de redes neurais artificiais apresentam resultados satisfatórias para extração de informações sintáticas e semântica em problemas de classificação de textos.

Objetivo de pesquisa:

Pesquisa para elaboração de modelos independentes de idioma e domínio textual baseados em redes neurais artificiais aplicado ao reconhecimento de características autorais à partir de documentos escritos nos idiomas português, inglês e espanhol, nos domínios de redes sociais, blogs, entrevistas, e devidamente anotados com informações de gênero, faixa etária, grau de religiosidade, grau de escolaridade, área de formação e visão política. Os córpus são: PAN-CLEF 2013, The Blog Authorship, B5-post, BRMoral e BlogSet-BR. Tais córpus serão utilizados para avaliação dos modelos de caracterização autoral propostos fazendo uso da medida F para comparação com sistemas de baseline.

Caracterização da solução em desenvolvimento:

Este estudo faz uso de técnicas de redes neurais artificiais para problemas de classificação de textos, tais como redes neurais de convolução (CNN) e redes neurais recorrentes (i.e., Long-Short Term Memory - LSTM). Além de métodos de representação distribuída de palavras (i.e., word embeddings) e caracteres (i.e., char embeddings), tais como word2vec e fasttext.

Fundamentos:
Os modelos de redes neurais do tipo CNN e LSTM, e representação distribuída de palavras, são conhecidos na literatura por capturar elementos semânticos e sintáticos da escrita.

Trabalhos relacionados:

Fatima apresenta em seu trabalho uma abordagem multilíngue para predição de gênero e idade, considerando características de estilo de escrita e conteúdo, além de informações de n-gramas de caracteres. Sap, desenvolve um modelo léxico independente de domínio de texto, considerando a predição de idade como um problema de regressão. Sierra, apresenta resultados iniciais com modelos baseados em word embeddings e redes neurais de convolução (CNNs) para predição de gênero e idade.

Validação:

As respostas dos modelo computacionais propostos e sistemas de baseline são comparadas, utilizando medida F e acurácia. Os córpus são separados entre conjuntos de treinamento e testes para tal avaliação.

Limitações, riscos e ameaças:

O estudo será limitado aos conjuntos de dados selecionados e às tarefas de caracterização autoral por eles suportadas.

Contribuição científica:

As contribuições científicas previstas para este trabalho são a definição de tarefas de caracterização autoral, assim como avançar a fronteira de conhecimento de abordagens baseadas em redes neurais artificiais para as tarefas de caracterização autoral de modo que gerem resultados de referência para futuros estudos desta área.

Contribuição tecnológica (se pertinente):
As contribuições tecnológica previstas para este trabalho são a organização de diversos córpus e tarefas de caracterização autoral, modelos baseados em redes neurais artificiais para estas tarefas e resultados de referência para futuros estudos desta área.

Método de pesquisa			
Gênero (escolha UM)	[ ] Pesquisa teórica	[x] Pesquisa prática	[ ] Pesquisa empírica [ ] Pesquisa metodológica
Natureza (escolha UMA)	[ ] Pesquisa básica	[x] Pesquisa aplicada	
Abordagem (escolha UMA)	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Revisão de literatura* (você pode escolher mais de uma)	[ ] Revisão narrativa [ ] Revisão descritiva [ ] Revisão de escopo	[ ] Meta-análise [ ] Revisão sistemática qualitativa [ ] Revisão *guarda-chuva*	[ ] Revisão teórica [ ] Revisão realística [ ] Revisão crítica
Procedimento técnico principal (escolha UM)	[x] Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto*	[ ] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação	[ ] Pesquisa etnográfica [ ] Teoria fundamentada em dados [ ] Ciência do projeto [ ] Outra Qual? _____
Análise de dados (você pode escolher mais de uma)	[x] Estatística descritiva [ ] Estatística inferencial	[ ] Teste estatístico [ ] Análise de conteúdo	[ ] Análise do discurso [ ] Outros: _____

* Definição de tipos de revisões de literatura estabelecida por 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

Próximas atividades:

Atividades para agosto/2019.

1. Revisão do texto (até 15 de agosto)
2. Depósito da dissertação (até 03 de setembro)
3. Defesa da dissertação (até 15 de setembro)

Opcional: Forneça um esquema gráfico que mostre aspectos de sua pesquisa. Por exemplo: um fluxograma para construção da sua solução ou um infográfico para sua proposta de pesquisa. Se necessário, use a quarta página.