V Workshop de Dissertações de Mestrado do PPgSI (2018)

FICHA DE PESQUISA

DADOS GERAIS				
Título do projeto de pesquisa	Caracterização autoral a partir de textos			
Orientando	Rafael Felipe Sandroni Dias			
Orientador(es)	Prof. Dr. Ivandré Paraboni			
Semestre no curso, na data do workshop	[ ] 2º semestre	[ ] 3º semestre	[X] 4º semestre	[ ] 5º semestre
Qualificação	[X] Qualificação já realizada em: 29/06/2018 [ ] Realização da qualificação planejada para: dd/mm/aaaa			
Defesa	Prazo máximo para depósito: 03/09/2019 Realização da defesa planejada para: 04/2019			
Linha e Área de pesquisa	Gestão e Desenvolvimento de Sistemas: [ ] BD [ ] Engenharia de Software [ ] Gestão de TI [ ] IHC		Inteligência de Sistemas: [X] IA [ ] Processamento Gráfico [ ] Reconhecimento de Padrões	
Área de aplicação	[ ] Ambientes corporativos [ ] Bioinformática [ ] Biometria [ ] Dispositivos móveis [ ] Educação	[ ] Educação a distância [ ] Governo eletrônico [ ] Internet [ ] Jogos [ ] Jogos sérios	[X] Língua Natural [ ] Linguística [ ] Processos de Negócio [ ] Quimioinformática	[ ] Redes Sociais [ ] Robótica [ ] Saúde [ ] [outro – escrever]
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.			

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	Na literatura, modelos de caracterização autoral (CA) utilizam, em sua grande maioria, conhecimentos linguísticos especializados para cada tipo de tarefa, idioma e domínio de texto. Estudos recentes de processamento de língua natural (PLN) têm obtido resultados promissores em modelos independentes de idioma e domínio com uso de técnicas baseadas em redes neurais artificiais.
Problema de pesquisa	Gerar uma abordagem independente do uso de conhecimento linguístico especializado para cada tarefa, idioma e domínio, para obter as características autoras de um indivíduo, e que apresentem maior eficiência que modelos tradicionais utilizados na área de CA, que fazem uso de conhecimento especializado para computar características (feature engineering) em textos.
Objetivo geral da pesquisa	Proposta de pesquisa para elaboração de um modelo independente de idioma e dominio 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 formulários do governo, e devidamente anotados com informações de gênero, idade, grau de religiosidade, grau de escolaridade, área de formação, visão política e localidade.
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.
Proposta para solução	Abordagem que permita o reconhecimento de características autorais de forma independente de conhecimento solução linguístico especializado para cada tarefa, idioma e domínio. Para isso, será desenvolvido uma estratégia combinando modelos de word embeddings e redes neurais profundas para extração de características e classificação de textos.
Dados	O projeto utiliza córpus públicos devidamente anotados com informações de gênero, idade, grau de religiosidade, grau de escolaridade, área de formação, visão política e localidade, nos idiomas português, inglês e espanhol. Tais córpus serão utilizados para avaliação do modelo de caracterização autoral independente de idioma e domínio. Os córpus são: PAN-CLEF 2013, PAN-CLEF 2014, The Blog Authorship, B5-Post, BRMoral, BlogSet-BR e e-sic.

Validação	Pretende-se avaliar se as respostas do modelo desenvolvido são superiores as respostas dos sistemas de baseline nas diversas tarefas de caracterização autoral em córpus de diversos idiomas e domínios de textos. Caso afirmativo, a abordagem proposta é tão boa quando os sistemas de baseline.
Limitações	O escopo deste projeto se limita ao estudo de técnicas de RNAs e modelos de word embeddings para classificação de textos, bem como resolver problemas de caracterização nos idiomas português, inglês e espanhol, e em domínios de textos de redes sociais, blogs e questionários.
Resultados esperados	Contribuições científicas: Este trabalho pretende avançar a fronteira de conhecimento sobre os problemas de caracterização autoral usando modelos independentes de domínio e idioma. Ao desenvolver modelos baseados em técnicas de redes neurais artificiais e word embeddings espera-se avançar o conhecimento da utilização de tais modelos na área de caracterização autoral, e em especial, o conhecimento de tarefas pouco exploradas na literaturapara o idioma português brasileiro.
	Contribuições tecnológicas: um modelo de caracterização autoral usando técnicas independentes de domínio e idioma;

MÉTODO DE PESQUISA				
Gênero	[ ] Pesquisa teórica	[X] Pesquisa prática	[ ] Pesquisa empírica	[ ] Pesquisa metodológica
Natureza	[ ] Pesquisa básica/pura	[X] Pesquisa aplicada		
Objetivo	[X] Pesquisa exploratória	[ ] Pesquisa descritiva	[ ] Pesquisa explicativa	[ ] [outro – escrever]
Abordagem	[X] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)	
Procedimento técnico principal	[X] Pesquisa experimental [X] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto* [ ] Pesquisa de levantamento	[X] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação [ ] Pesquisa etnográfica	[ ] Pesquisa netnográfica [ ] Teoria fundamentada em dados (*Grounded theory*) [ ] Ciência do projeto (*Design science*) [ ] [outro – escrever]	
Técnica(s) / Instrumento(s) de coleta de dados	[X] Medição [ ] Questionário [ ] Entrevista [ ] Grupos focais	[ ] Formulário [X] *Benchmark*	[ ] Observação (direta / participante) [ ] Diário de campo / notas de campo [ ] Análise documental (ou de artefatos) [ ] [outro – escrever]	
Técnica(s) de análise de dados	Análise quantitativa: [X] Estatística descritiva [ ] Estatística inferencial [ ] [outro – escrever]		Análise qualitativa: [ ] Análise de conteúdo [ ] Análise do discurso [ ] [outro – escrever]	

CRONOGRAMA

	2017												2018												2019											
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Revisão Bibliográfica							X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X											
Definição da proposta											X	X	X	X	X	X	X																			
PAN CLEF 2018																	X																			
Exame de Qualificação																		X																		
Organização dos córpus																		X	X	X																
Modelos computacionais																			X	X	X	X	X													
Avaliação																								X												
Refinamentos																									X											
Avaliação final																										X										
Análise dos resultados																											X									
Escrita da dissertação																					X	X	X	X	X	X	X	X								
Escrita de artigos																											X	X	X	X	X					
Depósito da dissertação																												X								