

Nome do Aluno: Samuel Naassom do Nascimento Porto

Nome do Orientador: Fábio José Ayres

Título: Estudo comparativo de modelos de linguagem natural baseados em aprendizado profundo.

Palavras Chave: *deep learning*, *natural language processing*, *attention models*, *transformers*, BERT, ELMo, GPT2.

1. Descrição do Problema e Revisão de Literatura

Grande parte da informação em nossa sociedade é armazenada e comunicada na forma de textos em linguagem natural. A área do conhecimento denominada “Processamento de Linguagem Natural” (em inglês: Natural Language Processing – NLP) refere-se ao conjunto de técnicas para a análise linguística por meio de ferramentas computacionais, com o objetivo de extrair informações relevantes de grandes volumes de documentos (LIDDY, E.D. 2001). Exemplos de uso de NLP incluem tradução de máquina, sumarização de texto, resposta a perguntas, recuperação de informação, modelagem por tópicos e mineração de opinião (CAMBRIA, E. 2014).

Em muitas aplicações de NLP é necessário o emprego de técnicas de aprendizado de máquina. Nos últimos dez anos o campo de aprendizado de máquina passou por uma mudança significativa com o advento das redes neurais profundas (HINTON G., OSINDERO S., TEH Y.W. 2006). Estas são redes neurais com um grande número de camadas escondidas e topologias inovadoras, que tem se mostrado bastante efetivas na construção de modelos preditivos. Particularmente, na área de processamento de imagens, o surgimento de arquiteturas de redes neurais profundas como AlexNet (KRIZHEVSKY A., SUTSKEVER I., HINTON .E. 2012), VGGNet (SIMONYAN K., ZISSERMAN A. 2015) e ResNet (HE K., ZHANG X., REN S., SUN J. 2016) trouxe um nível de desempenho de sistemas computacionais em atividades de entendimento de imagens que é compatível – e por vezes superiores – aos seres humanos.

Recentemente, novos modelos de linguagem natural em larga escala foram desenvolvidos, e existe a expectativa de que o uso destes modelos tenha impacto tão significativo em NLP como os modelos de imagem tiveram sobre o processamento de imagem e visão computacional. Pode-se citar como modelos de destaque na literatura científica recente o BERT (DEVLIN, J. *et al.* 2018), o ELMo (PETERS, M. E. *et al.* 2018) e o GPT (RADFORD, A. *et al.* 2018). Para exemplificar o avanço desse cenário, com base na representação de palavras por meio de vetores numéricos baseados nos contextos em que aparecem, o modelo BERT foi utilizado para fazer avaliação da similaridade semântica dos resultados de ensaios clínicos a fim de detectar a troca de resultados, com uma assertividade muito superior aos modelos de classificações convencionais (KOROLEVA A., KAMATH S., PAROUBEK P. 2019). Uma outra possível aplicação de NLP pode ser feita pelo uso do modelo ELMo para analisar as emoções das pessoas com base em seus tweets a fim de melhorar a respostas dos serviços de emergências em casos de calamidades naturais ou causadas pelo homem (SING N., ROY N., GANGOPADHYAY A. 2019).

2. Objetivo

Neste projeto serão avaliados os desempenhos dos modelos BERT, ELMo, e GPT2 em tarefas clássicas de processamento de linguagem natural, bem como será desenvolvida uma metodologia de

avaliação e customização de modelos para tarefas específicas, tais como classificação de texto em documentos jurídicos.

Os objetivos específicos deste projeto são:

- Conhecer técnicas de aprendizado profundo aplicadas a NLP;
- Comparar técnicas modernas de NLP;
- Desenvolver uma técnica de análise de textos jurídicos com aprendizado profundo.

3. Metodologia (Proposta)

O projeto deverá seguir as seguintes etapas:

a) Estudo acerca de redes neurais profundas e novas arquiteturas para NLP:
Embasamento computacional, matemático e estatístico com o intuito de saber quais as melhores aplicações de cada modelo para NLP.

b) Avaliação de desempenho dos modelos BERT, ELMo e GPT2:
A priori, os modelos que serão usados para realizar tarefas de NLP são BERT, ELMo e GPT2, contudo, com base na etapa anterior poderá ser selecionado algum outro modelo que também venha acrescentar na solução de uma dada tarefa.

Para cada modelo serão definidos:

- Tarefas e aplicações:** escolher entre sumarização, tradução, análise semântica, dentre outras tarefas. A escolha dependerá da etapa de estudo das técnicas e de suas potencialidades.
 - Métricas de desempenho:** Para cada tarefa a ser executada, a avaliação da performance do modelo pode ser mais conclusiva dependendo da métrica utilizada como pode se observar nas aplicações das métricas Cosine similarity (HUANG, A. 2008), Unilateral Weighted Jaccard similarity (SANTISTEBAN, J.; TEJADA-CÁRCAMO, J.), Perplexity (JELINEK, F. *et al.* 1997) and Word Error Rate (WANG, Y.-Y.; ACERO, A.; CHELBA, C. 2003).
 - Avaliação de desempenho:** Sabendo a métrica a ser utilizada o modelo é então avaliado e sua performance na tarefa é documentada. Existe uma miríade de datasets para avaliação de desempenho disponíveis na literatura, a depender da tarefa a ser avaliada. Nesta fase do projeto será escolhido o dataset apropriado.
- c) Desenvolvimento de metodologia de customização dos modelos para usos específicos
- d) Criação de uma base de treinamento, validação e teste usando documentos jurídicos
- e) Demonstração do uso de aprendizado profundo para entendimento de documentos jurídicos

4. Resultados Esperados

Como resultados desta pesquisa teremos uma análise do estado-da-arte em NLP, e o desenvolvimento de um modelo de aprendizado profundo para a análise de documentos jurídicos. Espera-se que este modelo seja útil a produção de insights a respeito de assuntos legais a partir de uma grande massa de decisões judiciais brasileiras.

5. Referências Bibliográficas

- LIDDY, E.D. Natural Language Processing. **Encyclopedia of Library and Information Science**, 2nd Ed. NY. Marcel Decker, Inc. 2001.
- CAMBRIA, E. Jumping NLP Curves: A Review of Natural Language Processing Research. **IEEE Computational intelligence magazine**. 2014.
- HINTON, G.; OSINDERO, S.; TEH, Y. W. A fast learning algorithm for deep belief nets. **Neural Computation**, Volume 18 Issue 7, Páginas 1527-1554, 2006.
- KRIZHEVSKY A., SUTSKEVER I., HINTON .E. Imagenet classification with deep convolutional neural networks. **Advances in Neural Information Processing Systems**, Páginas 1097-1105, 2012.
- SIMONYAN K., ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. **International Conference on Learning Representations**, 2015.
- HE K., ZHANG X., REN S., SUN J. Deep residual learning for image recognition. **Computer Vision and Pattern Recognition**, Páginas 770-778, 2016.
- DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **CoRR**, abs/1810.04805, 2018. <http://arxiv.org/abs/1810.04805>.
- PETERS, M. E. *et al.* Deep contextualized word representations. **arXiv preprint** arXiv:1802.05365. 2018.
- RADFORD, A. *et al.* Improving language understanding by generative pre-training. **URL preprint** Disponível em "https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf", Acesso em: 30/10/2018.
- KOROLEVA A., KAMATH S., PAROUBEK P. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. **Journal of Biomedical Informatics: X**, 2019.
- SING N., ROY N., GANGOPADHYAY A. Analyzing The Emotions of Crowd For Improving The Emergency Response Services. **Pervasive and Mobile Computing**, 2019.
- HUANG, A. Similarity measures for text document clustering. **Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)**, Christchurch, New Zealand, Páginas 9-56, 2008.
- SANTISTEBAN, J.; TEJADA-CÁRCAMO, J. Unilateral weighted Jaccard coefficient for NLP. **Fourteenth Mexican International Conference on Artificial Intelligence (MICA)**, IEEE, Páginas 14-20, 2015.
- JELINEK, F. *et al.* Perplexity—a measure of the difficulty of speech recognition tasks. **The Journal of the Acoustical Society of America**, v. 62, n. S1, Páginas S63-S63, 1977.
- WANG, Y.-Y.; ACERO, A.; CHELBA, C. Is word error rate a good indicator for spoken language understanding accuracy. **IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)**. IEEE, Páginas. 577-582, 2003.

Cronograma de atividades

	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Estudo de técnicas	■	■	■	■							
Avaliação de desempenho		■	■	■	■	■					
Metodologia customizada			■	■	■	■	■	■			
Criação base de dados			■	■	■	■	■	■	■		
Aplicação a docs. jurídicos			■	■	■	■	■	■	■	■	■
Atividades de escrita de relatório			■			■			■	■	■