Trabalho 4 PLN - Análise de Artigo
Discente: João Lucas Lima de Melo

Embeddings Jurídico: Representações Orientadas à Linguagem Jurídica Brasileira

ARTIGO

Quem são os autores? de quais universidades?

Fabrício A. do Carmo (Universidade Estadual do Maranhao), Ferdinando Serejo (Tribunal de Justiça do Estado do Maranhão), Antonio F. L. Jacob Junior (Universidade Estadual do Maranhao), Ewaldo E. C. Santana (Universidade Estadual do Maranhao), Fabio M. F. Lobato (Universidade Estadual do Maranhao, Universidade Federal do Oeste do Pará)

Onde foi publicado o artigo? é um veículo de renome na área?

O artigo foi publicado em 2023. Disponível nos Anais do XI Workshop de Computação Aplicada em Governo Eletrônico, publicado pela Sociedade Brasileira de Computação

TAREFA

Qual tarefa/problema está sendo abordado pelo trabalho? Tal tarefa é definida dentro do trabalho?

O trabalho possui como objetivo a construção de modelos de embeddings orientados ao âmbito jurídico para o português brasileiro.

Outros trabalhos são referenciados em que a tarefa é apresentada/definida? Quais outros autores são referenciados que abordam o mesmo ou problema semelhante?

O artigo menciona trabalhos relacionados a tarefa apresentada. É mencionado o trabalho de Smywinski-Pohl et al. (2019), que onde são treinados modelos Word2vec e Glove visando a criação de um dicionário que forneça uma interfaço entre palavras técnicas da justiça polonesa e palavras que possam ser entendidas por leigos. Chalkidis and Kampas (2019) também são mencionados pelo seu trabalho de criação de embeddings a partir de um grande corpus de dados jurídicos disponíveis na língua inglesa. Outros trabalhos relacionados a embeddings na língua portuguesa também são citados, como o de Cunha et al. (2022), cujo trabalho realizou treinamentos de modelos embeddings utilizando corpus da língua portuguesa, observando o impacto da parametrização dos modelos (e.g., dimensao do vetor), o tamanho do corpus de treinamento e do domínio.

A tarefa/fenômeno abordada é formalizada ou discutida informalmente?

A tarefa é formalizada, levantando os dados utilizados e os procedimentos a fim de construir os modelos de embeddings.

ABORDAGEM

Qual abordagem é proposta para solucionar a tarefa? A solução é baseada em que estratégias (aprendizagem de máquina, regras, etc.)? Como tais estratégias são empregadas (arquitetura, etc) para construir a solução proposta?

A abordagem propõe a criação de um corpus jurídico de treinamento, composto por dados pré-processados coletados de diferentes entidades jurídicas. Construído o corpus, os autores propõem o treinamento dos modelos de embeddings Word2vec e FastText levando em consideração diferentes configurações paramétricas como arquitetura, dimensão do vetor de características e épocas de treinamento. Para o aprendizado do modelo, os autores propõem validação cruzada do tipo *k-fold*, para $k = 5$. Os autores sugerem para a avaliação do modelo que o mesmo seja submetido à uma tarefa de classificação de petições iniciais fornecidas pelo Tribunal de Justiça do Estado do Maranhão. Para isso, é proposta a utilização de algoritmos de aprendiazem de máquina *K-Neighbors Neighbor* (KNN), Regressão Linear (RL), *Support Vector Machine* (SVM) e *Random Forest* (RF)

O que a literatura relacionada apresenta de diferente da abordagem dos autores? qual a contribuição original dos autores?

A inovação dos autores reside na aplicação dos métodos para a construção de modelos de embeddings no domínio jurídico para o português brasileiro, construindo uma codidifação mais eficiente para reconhecer relações contextuais entre palavras típicas da área.

METODOLOGIA

Quais dados são utilizados no artigo? Como foram obtidos? Estão disponíveis?

O corpus foi construído com dados públicos contendo textos de acórdãos do Tribunal Superior do Trabalho (TST)[1]; do Supremo Tribunal Federal (STF), por meio do Iudicium Textum Dataset (ITD) disponibilizado por Sousa and Del Fabro (2019); do Superior Tribunal Militar (STM)[2]; do Tribunal Superior Eleitoral (TSE)[3]; do Tribunal de Contas da União (TCU)[4]; de processos recorrentes disponibilizados pelo Conselho Nacional de Justic ̧a (CNJ) por meio do Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatorios (BNPR)[5]; e por meio de dados (e.g., decretos, acordãos e súmulas) disponíveis na plataforma LexML[6]. Com exceção do ITD e do TJMA, os dados foram coletados diretamente dos portais, via web crawlers.

Os autores apresentam experimentos? Como foram projetados?

Cada diferente instância de modelo treinado, Word2Vec e FastText levados em consideração suas arquiteturas (CBOW e Skip-gram) e dimensão do vetor de características (300 e 600) é instancializado como um modelo de embedding e é sujeito à avaliação. Os autores conduzem experimentos para os modelos treinados submetendo-os à quatro técnicas de aprendizagem de máquina para classificação, os já citados KNN, RL, RF, e SVM. As petições iniciais, fornecidas pelo TJMA, são os dados utilizados para a avaliação. As métricas coletadas são F1-macro e acurácia de cada um dos classificadores.

Os resultados são comparados com a literatura? como?

Os resultados não são comparados com a literatura. Há apenas comparação entre os diferentes modelos treinados e uma menção ao trabalho de Hartmann et al. (2017).

DISCUSSÃO

Os autores apresentam discussão dos resultados?

Os autores conduzem uma discussão dos resultados observados.

Os erros da abordagem são analisados qualitativamente?

Os autores se mantém na análise quantitativa dos resultados, sem desenvolver uma discussão mais minusciosa sobre o comportamento dos modelos e no que se traduz o valor de uma dada métrica em termos qualitativos.

Qual a sua impressão sobre o artigo?

A proposta do trabalho é justificada pela escassez dos modelos de embeddings aplicados ao domínio jurídico em português brasileiro, como exposto pelos autores. A abordagem sugerida tambem se mostra consistente, tanto a composição do corpus quanto a escolha dos modelos de embeddings a serem treinados. O método de treinamento, no entanto, não me soa suficientemente claro, ainda que seja descrito os fatores levados em consideração para a construção de cada instância de modelo treinado. A escolha dos classificadores e as métricas de avaliação me soam satisfatórias para perceber o desempenho de cada modelo treinado. No entanto, o trabalho carece de uma avaliação que não se restrinja a, exclusivamente, comparar os modelos treinados entre si, sem observar o que é proposto pela literatura.