14° Seminário de Iniciação Científica e Tecnológica

Classificação Automática de Texto para Despesas Públicas no Tribunal de Contas do Estado de Goiás usando Deep Learning

Jeferson Marques de Souza (IC), Daniel Xavier Sousa (PQ), Hugo Vinícius Leão e Silva (PQ)

PIBIC - Campus Anápolis
daniel.sousa@ifg.edu.br, hugo.vinicius@ifg.edu.br

***Palavras-Chave**: Aprendizado Profundo; Processamento de Linguagem Natural; Classificação de Despesas Públicas.*

Introdução

Despesas públicas devem ser fiscalizadas pelo Tribunal de Contas do Estado de Goiás (TCE-GO) para garantir a corretude dos empenhos realizados. Contudo, devido ao grande volume de dados se faz necessária uma abordagem de classificação automática para garantir que os empenhos estejam coerentes com os gastos realizados e evitar possíveis fraudes com a categorização indevida. Em parceria com o IFG, TCE-GO desenvolveu um projeto onde foi aplicado modelos de classificação para automatização desta tarefa. Contudo, o projeto havia aplicado somente modelos baseados em tecnologias *Emsemble* e SVM, sem explorar os avanços de *DeepLearning*. Considerando a complexidade da base de dados e a possibilidade de aplicação de conceitos como *transfer-learning* e *self-attention*, este trabalho explora novas aplicações em *DeepLearning* considerando principalmente abordagens do modelo BERT (DEVLIN *et al*, 2018). O nosso objetivo é verificar na literatura possíveis estratégias de execução do BERT que possam melhorar a acurácia dos modelos já aplicados no TCE-GO. Nossos experimentos mostram que melhoramos a acurácia de classificação em 17% para macro-F1 e 5% para micro-F1.

Metodologia

A intuição na aplicação dos modelos baseados no BERT é principalmente em melhorar a representação dos dados. Essa melhora ocorre por: i) utilizar dados já pré-processados em grandes volumes e ii) usar o contexto das palavras do problema para gerar uma representação mais discriminativa para a tarefa de classificação. Assim, utilizamos o BERT de duas maneiras: como modelo para classificação automática e exclusivamente para gerar uma nova representação dos dados. Na segunda maneira combinamos a nova representação com algoritmos tradicionais, como Florestas Aleatórias (FA) e SVM. Avaliações semelhantes já foram encontradas da literatura (BIRUNDA et al, 2021), mas nosso foco é sobretudo na análise da base de dados do TCE-GO, a inovação deste trabalho. Base esta que contém mais de 300 mil empenhos registrados e cerca de 460 classes distintas.

Resultados e Discussão

A Tabela 1 apresenta nossos resultados e o resumo entre diversos experimentos. Especificamente para a base de dados avaliada, percebemos melhor acurácia quando combinamos a representação gerada pelo BERT com a utilização de outros algoritmos de classificação. No caso, tivemos um aumento em 17% comparando a estratégia com o uso do BERT mais SVM (BERT-SVM), contra o uso do SVM com uma representação Bag-Of-Words somente. A análise pela métrica macro-F1 á bastante relevante em nosso caso, pois a base possui muitas classes e com alto grau de desbalanceamento.

Tabela 1. Experimentos realizados

Algoritmo/Modelo	Micro-F1	Macro-F1
SVM	86%	69%
BERT	89%	68%
BERT-FA	**91%**	80%
BERT-SVM	**91%**	**81%**

Conclusões

Esse trabalho apresenta os avanços obtidos na parceria entre IFG e TCE-GO com a aplicação de estratégias *DeepLearning*. Nós mostramos que uso da representação de dados com BERT, combinado com algoritmos tradicionais foram capazes de melhorar a acurácia da classificação automática de empenhos públicos aplicados na base de dados do TCE-GO. A combinação do BERT com SVM ou FA produziu resultados superiores aos algoritmos que exclusivamente usam Ensemble ou SVM. Em trabalhos futuros desejamos verificar se a mesma estratégia de execução apresenta bons resultados para outras tarefas, como é o caso da predição na regressão dos valores de empenhos públicos.

Agradecimentos

Agradecemos ao IFG por todo suporte e bolsa fornecida.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **CoRR**, 2018.

BIRUNDA, Selva et al. A Review on Word Embedding Techniques for Text Classification, **IDCTA**, 2021