XII JORNADA DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA
EVENTO ON-LINE
18 A 21 DE OUTUBRO
UNIVERSIDADE FEDERAL DA FRONTEIRA SUL

APLICAÇÃO DE MÉTODOS DE APRENDIZAGEM ATIVA NO CONTEXTO DE BUSCAS SISTEMÁTICAS NA LITERATURA DO COVID-19

JEAN CARLO HILGER[1], RAFAEL GAMA FERNANDES[2], GUILHERME DAL BIANCO[3]

1 Introdução

A Revisão Sistemática da Literatura (RSL) representa um dos passos iniciais para o desenvolvimento de pesquisas científicas. Na RSL, são identificados e analisados os principais trabalhos relacionados ao objeto de pesquisa e, como consequência, o estado-da-arte é revelado. No entanto, com o volume crescente de trabalhos científicos produzidos no mundo [Gu et.al., 2016], a busca por conteúdo relevante se torna uma tarefa cada vez mais desafiadora. Nesse contexto, métodos capazes de encontrar documentos relevantes, reduzindo o trabalho manual de identificar e descartar materiais irrelevantes, se tornam críticos para reduzir custos e tempo de desenvolvimento de pesquisas científicas.

Abordagens supervisionadas para a classificação de textos são utilizadas para classificar documentos como “relevantes” ou “irrelevantes” [Sebastiani, 2002]. Na classificação supervisionada, um conjunto pré-selecionado de treinamento, já rotulado pelo usuário como relevante ou não, é fornecido para o algoritmos de classificação, que irá aprender os padrões ou comportamentos da base de dados. Tal aprendizado é aplicado em documentos não rotulados para capturar os documentos que apresentam padrões similares ao treinamento [Sebastian et al., 2012]. Em [Cormack et. al., 2018], por exemplo, é utilizado um algoritmo supervisionado para produzir um ordenamento de documentos de acordo com a relevância do termo de consulta para a RSL. O método, chamado de KNEE iterativamente aprende os padrões presentes em documentos relevantes ou não-relevantes informados pelo usuário. No entanto, uma das limitações da aprendizagem supervisionada é como construir o conjunto de treinamento manualmente sem um conhecimento prévio da base de dados. Note que se o treinamento não contiver determinados padrões presentes em documento relevante,

1 Acadêmico de Ciência da Computação, UFFS, *campus* Chapecó, **Bolsista**, contato: hilgerjeancarlo@gmail.com
2 Acadêmico de Ciência da Computação, UFFS, *campus* Chapecó, contato: rafagama@outlook.com.br
3 Doutor,UFFS, **Orientador**.

dificilmente o algoritmo de classificação irá encontrar tal conjunto de documentos, podendo resultar em um alto número de documentos relevantes não identificados [Bianco et al., 2015].

No entanto, a lacuna de como extrair (mapear) características informativas de documentos não estruturados ainda permanece uma vez que os métodos demandam dados numéricos (técnicas de aprendizado de máquina não operam sobre dados textuais). Técnicas de mapeamento de palavras para números, como TF-IDF, produzem um número elevado de termos (milhões de palavras), aumentando o custo do processo e reduzindo a qualidade. Além disso, este modelo não leva em consideração a ordem dos termos e suas relações, ou seja, o sentido das frases é perdido ou ignorado. Neste contexto, criou-se a motivação para aplicar técnicas de aprendizado profundo (*deep learning*) em dados textuais. O objetivo da técnica BERT (Bidirectional Encoder Representations from Transformers), por exemplo, é de generalizar informações textuais extraindo padrões de similaridade e semântica dos documentos [Vaswani et.al, 2017]. BERT utiliza *autoencoders* para aprender, na fase de pré-treino, e outro na fase de ajuste para a calibração em relação à tarefa em específico. De forma simplista, na fase de pré-treino é inserido como entrada sentenças, na qual algumas palavras são omitidas, e o objetivo é reconstruir a sentença a partir do seu contexto. O treinamento visa "ensinar" o modelo a reduzir erros e se tornar capaz de identificar os termos semelhantes na sentença[Zhu et al. 2015].

2 Objetivos

Analisar experimentalmente abordagens para a extração de features com base em técnicas de *deep learning*, no contexto do COVID-19, para a revisão sistemática da literatura.

3 Metodologia

A partir de uma análise bibliográfica compreendeu-se as principais alternativas para resolução do problema proposto. A confecção da solução sucedeu-se, inicialmente, utilizando um modelo BERT pré-treinado como extrator de características (os textos originais, inseridos como entrada do modelo, resultam em vetores numéricos, ditos *embeddings*). Além do formato descrito, variou-se características arquiteturais do modelo BERT. Aplicou-se também, um modelo SBERT, para extração de características. Para este, operou-se com um formato de calibração iterativa: a cada iteração do algoritmo base, o modelo extrator de características é re-treinado com os textos classificados pelo usuário, buscando uma precisão - a nível de características - cada vez maior.

4 Resultados e Discussão

Nesta seção, serão apresentados os resultados encontrados a partir da utilização de técnicas baseadas em *autoencoders*: BERT e SBERT. BERT busca apreender relações entre palavras em documentos textuais. Foram usados modelos pré-treinados em grandes coleções de dados para extrair os padrões e coocorrência de palavras ou termos. Como método base, foi utilizado o KNEE-TFIDF [Cormack et. al., 2015] na qual as características são extraídas com base na estratégia de ponderação de termos baseado na sua frequência (TF-IDF). Após, o algoritmo SVM é aplicado para identificar os documentos relevantes. Já o método proposto KNEE-BERT utiliza o método BERT como extrator de features com intuito de extrair as relaçẽos entre as palavras. Como o modelo utilizado BERT é pré-treinado, um conjunto reduzido de características é produzido (por exemplo, 768 características). Por fim, no KNEE-SBERT, as características são extraídas a partir do SBERT e como classificador são computadas similaridade *coseno* entre as características extraídas.

4.1 Bases de dados e Métricas

A base de dados utilizada neste trabalho é baseada em artigos científicos com o tópico do COVID-19. A base é oriunda do desafio chamado TREC-COVID[4] criado para avaliar métodos de recuperação de informação. Os artigos são manualmente avaliados como relevantes de acordo com 50 consultas previamente definidas. Neste trabalho, devido a restrição de espaço, somente 10 consultas (ou tópicos) foram utilizados. Para avaliação da eficácia dos métodos, foi utilizada a métrica de revocação (*Recall*), na qual, é computado a razão entre o número de documentos relevantes recuperados e o total de relevantes. A métrica de área da curva (*Area Under the Curve* - AUC) computa o ganho em relação a área da curva de recall (eixo y) e número de documentos rotulados (eixo x), com objetivo de permitir a comparação numérica entre os métodos. Quanto maior o valor da área, melhor é o desempenho do método.

4.2 Resultado Experimental

Os gráficos da Figuras 1 ilustram a curva de *recall* em relação ao custo de rotulação (documentos rotulados manualmente) em 2 tópicos usando as abordagens KNEE-SBERT, KNEE-BERT, KNEE-TFIDF. Quanto maior for a inclinação da curva melhor é a capacidade do método em recuperar documentos relevantes (informativos para o usuário). Conforme pode ser observado, o método KNEE-SBERT apresenta um leve ganho sobre os demais na

4Maiores informações podem ser encontradas em https://ir.nist.gov/trec-covid/.

maioria dos casos. Tal ganho pode ser melhor visualizado a partir da análise da Tabela 1, na qual é computado a área ocupada pela curva. Quanto maior a área, melhor é a capacidade do método em recuperar documentos relevantes. Dos 10 tópicos analisados, a abordagem KNEE-SBERT apresenta uma área superior em 8 casos se comparado a abordagem KNEE-TFIDF, chegando a um ganho de 12% no tópico 9. Já as áreas do método KNEE-SBERT foram melhores que todas as áreas do KNEE-BERT, chegando a um ganho de 24%. A partir desse experimento foi possível avaliar a promissora capacidade do método KNEE-SBERT em identificar um número reduzido de características (se comparado ao método TF-IDF) com ganho de desempenho se comparado as técnicas analisadas. O conjunto de características extraído pelo KNEE-SBERT se mostra mais informativo e capaz de melhorar o desempenho de classificação.

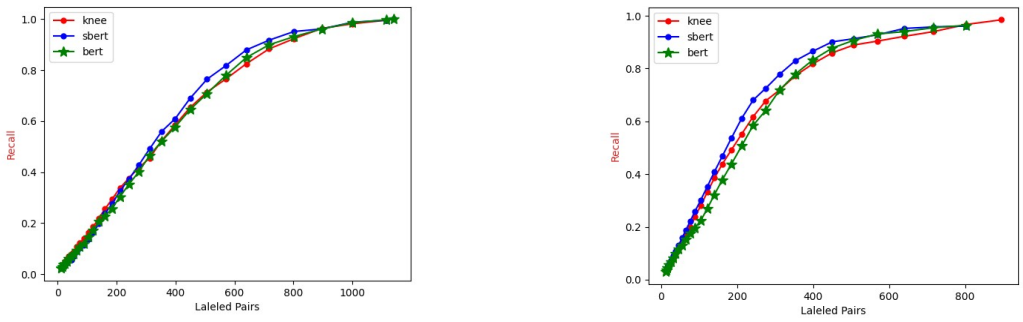

Figura 1: Gráficos ilustrando o Recall e o número de documentos rotulados. Quanto maior a área da curva melhor é o resultado da abordagem.

Tabela 1: Comparação da variação da área da curva dos métodos KNEE-SBERT vs KNEE-TFIDF e KNEE-BERT. Quanto maior (mais positivo for a variação) maior é o ganho.

Tópico	KNEE-SBERT vs KNEE-TFIDF	KNEE-SBERT vs KNEE-BERT
1	+2,84	+7,97
2	+5,94	+9,35
3	-2,06	+3,42
4	**+6,39**	**+11,87**
5	+1,19	+8,3
6	+1,65	+5,85
7	-3,06	+5,31
8	+1,59	+10,7
9	**+12,13**	**+24,28**
10	+0,67	+2,49

5 Conclusão

Este trabalho teve como objetivo avaliar experimentalmente a inserção de *autoencoders* para a extração de features no contexto de buscas sistemáticas da literatura do Covid-19. Os experimentos executados demonstraram resultados promissores em utilizar o método SBERT como extrator de características além de reduzir o número de características geradas. Pretende-se, em trabalhos futuros, avaliar experimentalmente os métodos aqui estudados em outras bases de dados para analisar os resultados.

Referências Bibliográficas

Gu, Xin, and Karen L. Blackmore. **Recent trends in academic journal growth.** Scientometrics 108.2 (2016).

Sebastiani, Fabrizio. **Machine learning in automated text categorization.** ACM computing surveys (CSUR) 34.1 (2002).

Yu, Zhe, et al. **Improving Vulnerability Inspection Efficiency Using Active Learning**. arXiv preprint arXiv:1803.06545 (2018).

Dal Bianco, Guilherme, et al. **A practical and effective sampling selection strategy for large scale deduplication.** IEEE TKDE 27.9 (2015).

Vaswani, Ashish, et al. "**Attention is all you need.**" In: *NeurIPS* (2017).

Yukun Zhu, et. al.. 2015. **Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.** In arXiv arXiv:1506.06724.

Cormack, Gordon V., and Maura R. Grossman. "**Scalability of continuous active learning for reliable high-recall text classification.**" Proceedings of CIKM. ACM, 2016.

Palavras-chave: Aprendizagem ativa, Revisão sistemática da Literatura, COVID-19

Nº de Registro no sistema Prisma: PES-2021-0140

Financiamento: CNPq