Verificação de Factualidade

Membro: João Carlos Cerqueira (jc.cerqueira13@gmail.com)

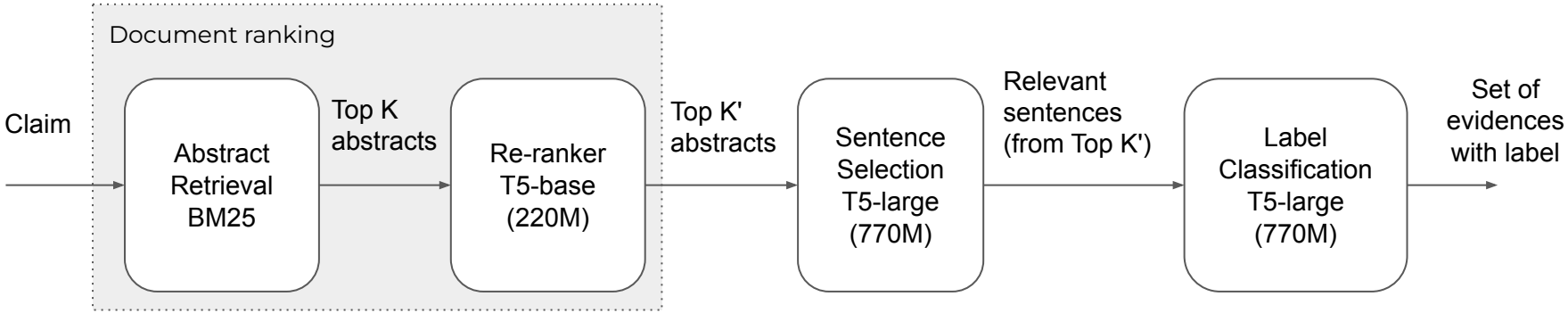
01 de dezembro de 2022

Contexto

- O objetivo do projeto é implementar um sistema capaz de **automatizar a verificação da veracidade de afirmações** (verdadeiras ou falsas), além de classificar as afirmações, o sistema deve **fornecer evidências**;
- O sistema será formado por 3 estágios: (1) ranking de documentos relevantes, (2) seleção de sentenças importantes e (3) classificação;
- Serão utilizados múltiplos modelos T5 (3 modelos), como foi feito em [1];
- O dataset a ser utilizado é o SciFact [2], composto de 1.400 afirmações da literatura científica médica e biomédica

Arquitetura

Mudança: substituímos dois modelos T5-Base (220M) por modelos T5-Large (770M) devido aos resultados parciais observados. Ponto importante: no artigo de referência, todos os modelos T5 foram implementados utilizando-se o T5-3B, com 3 bilhões de parâmetros.



Resultados – Abstract Retrieval & Re-ranking

Os sistemas de busca/ranking de *abstracts* relevantes foram avaliados a partir da métrica de recall, com a restrição de apenas os *top_k* serem contabilizados para a métrica, onde *k=3*.

Tabela: Comparação do recall para a tarefa de ranking de documentos, limitados a até 3 (R@3) e 5 (R@5) documentos selecionados, dataset de desenvolvimento.

Modelo	R@3	R@5
Oracle	97,61%	100,00%
TF-IDF	69,38%	75,60%
BM25	79,90%	84,69%
T5 (MS MARCO)	86,12%	**89,95%**
T5 (MS MARCO MED)	85,65%	89,00%
T5 (SciFact)	**86,60%**	89,40%
T5 (nossa solução)	85,65%	88,52%

O desempenho foi semelhante para todas as implementações, apesar da grande diferença entre as soluções (paradigmas diferentes, tamanho de modelos diferentes).

Resultados – Sentence Selection

A tarefa de Sentence Selection foi avaliada em relação a precisão, recall e F1-Score para a predição da classe positiva (isto é, sentença importante). Os resultados que obtivemos estão em linha com o artigo de referência.

Tabela: Desempenho da tarefa de seleção de sentenças importantes, dataset de desenvolvimento.

Modelo	Precisão	Recall	F1-Score
RoBERTa-large	73,71%	70,49%	72,07%
T5 (3-B)	**79,29%**	73,22%	76,14%
T5 (nossa solução)	76,86%	**76,23%**	**76,54%**

Detalhe importante: como os exemplos negativos do dataset de treino e desenvolvimento foram criados artificialmente, é possível que exista divergências na metodologia e os resultados não sejam diretamente comparáveis.

Resultados – Label Classification (T5-Small)

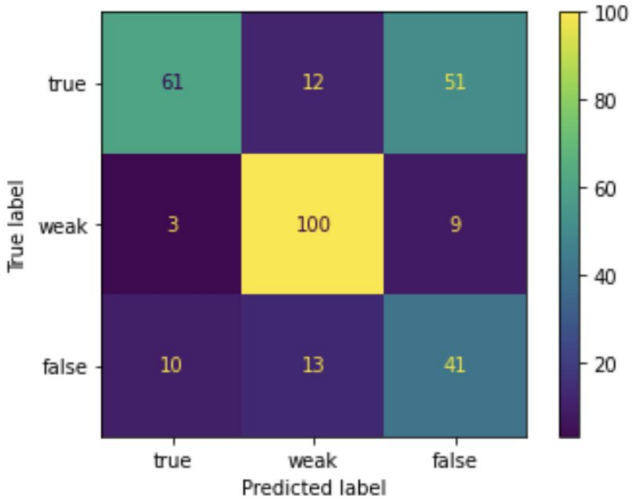
À direita temos os resultados do treinamento do **T5-Small** na da etapa de ***Label Classification*** para o dataset de **desenvolvimento**.

	precision	recall	f1-score	support
true	0.82	0.49	0.62	124
weak	0.80	0.89	0.84	112
false	0.41	0.64	0.50	64

Referência [1]:

Method	Label	P	R	F1
RoBERTa-large	SUPPORTS	92.56	81.16	86.49
	NOINFO	74.82	**92.86**	82.87
	REFUTES	77.05	66.20	71.21
T5	SUPPORTS	**93.13**	**88.41**	**90.71**
	NOINFO	**85.25**	**92.86**	**88.89**
	REFUTES	**86.76**	**83.10**	**84.89**

Table 8: Comparison of different label prediction models, on SCIFACT's development set.



Resultados – Label Classification (T5-Large)

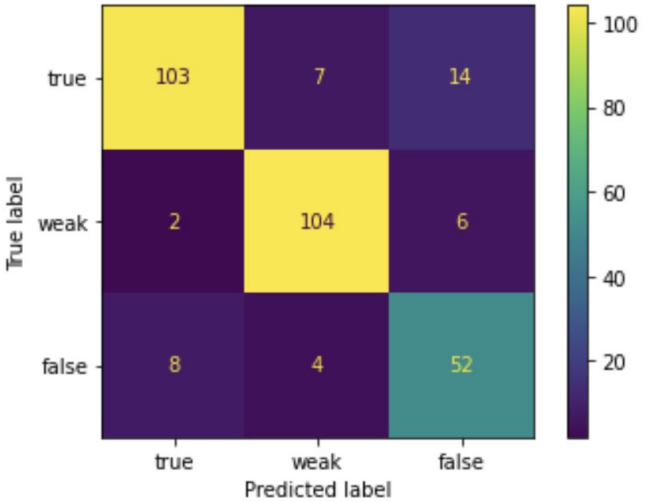
À direita temos os resultados do treinamento do **T5-Large** na da etapa de ***Label Classification*** para o dataset de **desenvolvimento**.

	precision	recall	f1-score	support
true	0.91	0.83	0.87	124
weak	0.90	0.93	0.92	112
false	0.72	0.81	0.76	64

Referência [1]:

Method	Label	P	R	F1
RoBERTa-large	SUPPORTS	92.56	81.16	86.49
	NOINFO	74.82	**92.86**	82.87
	REFUTES	77.05	66.20	71.21
T5	SUPPORTS	**93.13**	**88.41**	**90.71**
	NOINFO	**85.25**	**92.86**	**88.89**
	REFUTES	**86.76**	**83.10**	**84.89**

Table 8: Comparison of different label prediction models, on SCIFACT's development set.



Resultados – Label Classification (Comparativo final)

O desempenho da solução implementada é inferior ao apresentado na referência [1]. Hipótese: a diferença de desempenho é devido à diferença do número de parâmetros (770 milhões vs. 3 bilhões). Como referência, o modelo RoBERTa-large tem 354 milhões de parâmetros.

Tabela: Desempenho para a tarefa de classificação, dataset de desenvolvimento.

Modelo	Classe	Precisão	Recall	F1-Score
RoBERTa-large	SUPPORTS	92,56%	81,16%	86,49%
	NOINFO	74,82%	**92,86%**	82,87%
	REFUTES	77,05%	66,20%	71,21%
T5 (3-B)	SUPPORTS	**93,13%**	**88,41%**	**90,71%**
	NOINFO	85,25%	**92,86%**	**88,89%**
	REFUTES	**86,76%**	**83,10%**	**84,89%**
T5 (nossa solução)	SUPPORTS	88,62%	87,90%	88,26%
	NOINFO	**86,54%**	80,36%	83,33%
	REFUTES	69,86%	79,69%	74,45%

Resultados final – Pipeline completo (oráculo)

Modelo	Precision	Recall	F1-Score
Label Only			
Oracle (referência, 3B)	**92,70%**	**78,95%**	**85,27%**
Oracle (nosso)	59,09%	37,32%	45,75%
Label + Rationale			
Oracle (referência, 3B)	**88,76%**	**75,60%**	**81,65%**
Oracle (nosso)	49,24%	31,10%	38,12%
Selection Only			
Oracle (referência, 3B)	**83,54%**	**72,13%**	**77,42%**
Oracle (nosso)	64,82%	44,81%	52,99%
Selection + Label			
Oracle (referência, 3B)	**78,16%**	**67,49%**	**72,43%**
Oracle (nosso)	37,15%	25,68%	30,37%

Resultados final – Pipeline completo

Modelo	Precision	Recall	F1-Score
	Label Only		
T5-3B (referência)	**65,07%**	**65,07%**	**65,07%**
T5-Large (nosso)	37,00%	35,41%	36,19%
	Label + Rationale		
T5-3B (referência)	**61,72%**	**61,72%**	**61,72%**
T5-Large (nosso)	30,50%	29,19%	29,83%
	Selection Only		
T5-3B (referência)	**64,81%**	**57,37%**	**60,87%**
T5-Large (nosso)	42,18%	41,26%	41,71%
	Selection + Label		
T5-3B (referência)	**60,80%**	**53,83%**	**57,10%**
T5-Large (nosso)	25,14%	24,59%	24,86%

Conclusão

→ Foi possível observar o impacto do tamanho dos modelos de linguagem natural versus o desempenho em cada uma das etapas do fluxo de verificação de factualidade: tamanho do modelo é relevante para a tarefa de Classificação, mas influenciou pouco no resultado das outras etapas.

→ É interessante investigar maneiras de melhorar o desempenho do modelo desenvolvido frente aos resultados obtidos no artigo de referência:

- A **Classificação** é fortemente influenciada pelo desbalanceamento de classes. É possível corrigir o desempenho utilizando-se oversampling, ou então fazendo o rebalanceamento dos logits dos tokens obtidos na saída do modelo.
- O valor de corte dos logits da etapa de **Seleção de Sentenças** impacta muito o desempenho, é possível buscar por valores ótimos desse parâmetro.
- Revisar a função de avaliação de desempenho.

Referências

[1] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Scientific Claim Verification with VERT5ERINI. arXiv:2010.11930, October 2020. arXiv: 2010.11930. http://arxiv.org/abs/2010.11930

[2] Homepage about the dataset SciFact, provided by Allen Institute for AI: https://leaderboard.allenai.org/scifact/submissions/get-started

[3] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or Fiction: Verifying Scientific Claims. arXiv:2004.14974, October 2020, arXiv: 2004.14974. http://arxiv.org/abs/2004.14974

[4] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. https://microsoft.github.io/msmarco/

[5] Wikipedia: Okapi BM25 (https://en.wikipedia.org/wiki/Okapi_BM25)

[6] SciFact scoring function implementation details (with examples) https://github.com/allenai/scifact/blob/master/doc/evaluation.md

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910.10683.

[8] Pyserini, a Python toolkit for reproducible information retrieval research with sparse and dense representations. https://github.com/castorini/pyserini

[9] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In Findings of EMNLP.

[10] Hugging Face, MonoT5, a T5-base reranker fine-tuned on the MS MARCO passage dataset for 100k steps (or 10 epochs). https://huggingface.co/castorini/monot5-base-msmarco.

[11] MS MARCO, a collection of datasets focused on deep learning in search. https://microsoft.github.io/msmarco

[12] PaLM: Scaling Language Modeling with Pathways https://arxiv.org/abs/2204.02311