Detecção de Depressão nas Mídias Sociais usando Transformers com Aprendizado Federado

Arthur B. Vasconcelos[1], **Rafaela Brum**[1], **Aline Paes**[1], **Lúcia Maria de Assumpção Drummond**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF) – Niterói, RJ – Brasil

{athurbittencourt, rafaelabrum}@id.uff.br, {alinepaes, lucia}@ic.uff.br

Abstract. *Many initiatives investigate the automatic detection of depression through publications on social media, employing machine learning models. The most adequate way to obtain datasets would be through users consenting to share their data. Considering the sensibility of the data, it is not recommeded to share these publications to other research groups. A solution to this issue is to employ Federated Learning techniques to exchange the trained models instead of datasets. In this work we have investigated how to combine trained Transformers, the state of the art architecture for word embedding, to tackle the automatic detection of depression from social media.*

Resumo. *Diversas iniciativas investigam a detecção automática de depressão através de publicações em redes sociais utilizando de modelos de aprendizado de máquina. A forma mais adequada de obter conjuntos de dados seria pelo consentimento dos usuários em compartilhar seus dados. Porém, compartilhar estas publicações para outros grupos de pesquisa nem sempre é recomendado dada a sensibilidade dos dados. Uma solução para este problema, é se valer de técnicas de Aprendizado Federado para trocar modelos no lugar de dados. Neste trabalho, investigamos como combinar Transformers, a arquitetura estado da arte para word embedding, treinados para abordar a detecção automática de depressão nas mídias sociais.*

1. Introdução

A depressão tem acometido um número cada vez maior de pessoas ao redor do mundo [World Health Organization 2017]. Entretanto, muitos não conseguem obter a ajuda adequada. Assim, diversas iniciativas têm investigado a detecção automática de depressão baseada em seus sintomas a partir das publicações de usuários de redes sociais, empregando modelos de aprendizado de máquina.

A forma mais adequada de obter dados para o treinamento destes modelos é por consentimento explícito dos usuários. Pois considerando usuários que se voluntariem para participar do estudo, é possível obter a rotulação das publicações e dos usuários por meio de inventários de depressão, como o inventário de depressão de Beck (BDI) [Beck et al. 1961], que são questionários vastamente empregados e estudados na área da psicologia. Entretanto, o volume de publicações costuma ser muito menor do que uma coleta aberta a partir das plataformas de redes sociais, uma vez que é necessário que indivíduos se voluntariem para participar do estudo. Para aumentar o volume de dados, diversos grupos de pesquisa que trabalham com o mesmo tema poderiam compartilhar suas

bases de dados. Porém, considerando a sensibilidade dos dados, nem sempre é possível ou recomendável compartilhar as publicações fornecidas pelos usuários para que outros pesquisadores possam treinar seus modelos.

Uma possível solução para o problema do compartilhamento dos dados é se valer de técnicas de Aprendizado Federado [McMahan et al. 2017], em que o aprendizado se dá por meio de uma federação de clientes, coordenados por um servidor central. Os clientes rodam seu aprendizado localmente, em cima de seus próprios dados, e comunicam ao servidor o resultado deste treinamento. Desta forma, os clientes não precisam trocar seus dados de treinamento entre si, mas apenas o modelo para que o servidor possa agregá-lo aos demais modelos. A maioria das redes sociais são baseadas em textos, e o estado da arte para aprender modelos de aprendizado de máquina a partir de textos faz uso da arquitetura de Transformers [Vaswani et al. 2017].

2. Metodologia e Resultados Experimentais

A base de dados considerada neste trabalho é a eRisk2021 [Parapar et al. 2021], que contém publicações do Reddit anotadas com a classe depressiva ou não (classificação binária). Foram considerados três modelos BERT [Devlin et al. 2019] para o ajuste fino usando a biblioteca transformers do HuggingFace [Wolf T. et al. 2020]. Os modelos foram treinados utilizando o otimizador AdamW, com taxa de aprendizado de 1e-5. O ambiente de execução usado foi a máquina DGX da NVidia, equipada com um processador Xeon E5-2698 V4. e 8 GPUs Tesla P100 SXM2.

Inicialmente, o modelo base fornecido pela biblioteca foi executado sem nenhum ajuste nos parâmetros. A seguir, este mesmo modelo passou pelo processo de *fine-tuning* de forma centralizada por 50 épocas em uma única GPU. O modelo baseline também foi o ponto de partida para o *fine-tuning* de forma federada, com cinco clientes sendo treinados por uma época por 50 rounds, usando o *framework* Flower [Beutel et al. 2020], e tendo FedOpt como estratégia de agregação. O dataset foi distribuído de forma homogênea entre os clientes, isto é, ele foi separado de forma aleatória, mas preservando a proporção de exemplos entre as duas classes. Os cinco clientes e servidor compartilharam três GPUs. O terceiro modelo foi treinado da mesma forma que o segundo, mas rodando por 10 rounds, com cada cliente treinando cinco épocas cada. Foi usada a mesma partição do conjunto de dados, e a mesma organização de GPUs.

A Tabela 1 apresenta os resultados da classificação por publicação e por usuário (ou seja, considerando todas as suas publicações) dos três modelos, em comparação com o Baseline. Os resultados por publicação foram metrificados por acurácia, precisão, recall, f1 score, loss e tempo de treinamento em horas. Os resultados por usuários foram metrificados por precisão (u_p), recall (u_r) e f1 score (u_{f1}).

Em todos os casos, foi possível observar que o modelo piora a cada época do treinamento, em um fenômeno similar ao de *overfitting*. Os resultados da avaliação do modelo federado de 50 rounds são melhores do que os do centralizado, enquanto o modelo de 10 rounds teve os piores resultados. Também, podemos observar que os resultados por usuário não mudaram muito entre os modelos treinados. Pelos testes preliminares realizados, foi possível verificar que isso decorre principalmente pelo treinamento considerar apenas as postagens individualmente. Assim, o modelo acaba por avaliar a maioria dos exemplos como depressivos, mesmo quando o usuário não é.

Table 1. Resultados por postagem e usuário

Modelo	accur.	prec.	recall	f1 score	loss	u_p	u_r	u_{f1}	tempo(hr)
Baseline	0.2582	0.8276	0.0210	0.0409	**0.1282**	0.1250	**0.5000**	0.2000	0.0000
Centralizado	0.6054	0.7431	0.7290	0.7360	0.5058	**0.3750**	**0.5000**	**0.4286**	**6.2485**
Federado (50 rounds)	**0.6142**	0.7444	**0.7440**	**0.7442**	0.4245	**0.3750**	**0.5000**	**0.4286**	8.9280
Federado (10 rounds)	0.5880	**0.7450**	0.6898	0.7164	0.4621	0.3667	0.4583	0.4074	7.6249

3. Conclusões e Trabalhos futuros

O Modelo Federado de 50 rounds apresentou melhores resultados do que o modelo centralizado, o que pode ser explicado pela tendência deste arranjo de desacelerar o aprendizado entre os clientes. Ainda assim, a maior parte dos usuários foi classificada como depressivo, mesmo quando não apresentam sinais, o que indica que o uso exclusivo do BERT neste trabalho não é adequado. Como trabalho futuro, os próximos passos serão usar outro pré-treinamento, diferente do checkpoint bert-base-uncased, e utilizar outras arquiteturas de processamento de linguagem.

References

Beck et al. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571.

Beutel et al. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Devlin et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

McMahan et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Parapar et al. (2021). Overview of erisk 2021: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, page 324–344, Berlin, Heidelberg. Springer-Verlag.

Vaswani et al. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wolf T. et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization (2017). Technical report, Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO .