Título: Automatic Misinformation Detection About COVID-19 in Brazilian Portuguese WhatsApp Messages

Data: 30/11/2023

Horário: 10h00

Local: Videoconferência

Resumo:

Durante a pandemia da COVID-19, o problema da desinformação voltou a surgir, de forma

bastante intensa, através das redes sociais. Em muitos países em desenvolvimento, como o Brasil, uma das principais fontes de desinformação é o aplicativo de mensagens WhatsApp. No entanto, devido à natureza das mensagens privadas do WhatsApp, ainda existem poucos métodos de detecção de desinformação desenvolvidos especificamente para esta plataforma. Além disso, um modelo de detecção de desinformação construído para Twitter ou Facebook pode ter um desempenho ruim quando usado para classificar mensagens do WhatsApp. Nesse contexto, a detecção automática de desinformação (MID) sobre o COVID-19 em mensagens do WhatsApp em português do Brasil torna-se um desafio crucial. Neste trabalho, apresentamos o COVID-19.BR, um conjunto de dados de mensagens do WhatsApp sobre coronavírus em português brasileiro, coletadas de grupos públicos brasileiros e rotuladas manualmente. Além disso, avaliamos uma série de classificadores de desinformação combinando diferentes técnicas. Nosso melhor resultado utilizando aprendizado de máquina clássico alcançou F1 de 0,799, e a análise dos erros indica que eles ocorrem principalmente pela predominância de textos curtos. Quando são filtrados textos com menos de 50 palavras, a pontuação F1 sobe para 0,866. Além disso, propomos uma nova abordagem, chamada MIDeepBR, baseada em redes neurais BiLSTM, operações de pooling e mecanismo de atenção, que é capaz de detectar automaticamente desinformação em mensagens do WhatsApp em português brasileiro. MIDeepBR supera as abordagens clássicas de aprendizado de máquina, alcançando F1 de 0,834. Finalmente, exploramos um método de interpretabilidade post-hoc chamado LIME para explicar as previsões das abordagens de MID. Além disso, aplicamos uma ferramenta de análise textual chamada LIWC para analisar as características linguísticas das mensagens do WhatsApp e identificar aspectos psicológicos presentes em mensagens com desinformação e com não desinformação. Os resultados indicam que é viável compreender aspectos relevantes das previsões do modelo de MID e encontrar padrões nas mensagens do WhatsApp sobre o COVID19. Assim, esperamos que estas descobertas ajudem a compreender os fenômenos de desinformação sobre a COVID-19 nas mensagens do WhatsApp.

Banca examinadora:

- Prof. Dr. Javam de Castro Machado (MDCC/UFC - Orientador)
- Prof. Dr. José Maria da Silva Monteiro Filho (MDCC/UFC - Coorientador)
- Prof. Dr. César Lincoln Cavalcante Mattos (MDDC/UFC)
- Prof. Dr. Sergio Lifschitz (PUC-RIO)