Telegram4DS: Um Conjunto de Dados de Perguntas de Grupos Brasileiros do Telegram sobre Ciência de Dados e IA

Leonardo Gargano[1], Adriana S. Vivacqua[1]

[1] Programa de Pós-graduação em Informática, Instituto de Computação – Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro – RJ – Brasil

leonardo_gargano@ufrj.br, avivacqua@ic.ufrj.br

Resumo. *Aplicativos de mensagens instantâneas concentram discussões técnicas que raramente são capturadas por plataformas tradicionais de perguntas e respostas. Este trabalho apresenta o Telegram4DS, um conjunto de dados rotulado manualmente com 2.000 perguntas coletadas em 10 grupos públicos do Telegram focados em Ciência de Dados e Inteligência Artificial no Brasil. A partir de 631.014 mensagens brutas, foram selecionadas perguntas aleatoriamente e aplicada análise temática para classificá-las em quatro categorias: mercado, cursos, dúvidas gerais e material.*

Abstract. *Instant messaging applications host technical discussions that are rarely captured by traditional question-and-answer platforms. This paper presents Telegram4DS, a manually labeled dataset containing 2,000 questions collected from 10 public Telegram groups focused on Data Science and Artificial Intelligence in Brazil. From a corpus of 631,014 raw messages, questions were randomly selected and subjected to thematic analysis to classify them into four categories: job market, courses, general inquiries, and materials.*

1. Introdução

As plataformas de mensagens, especialmente aquelas voltadas para dispositivos móveis, têm se tornado cada vez mais presentes na sociedade contemporânea [Baumgartner et. al., 2020]. Atualmente, é comum observar uma grande quantidade de usuários ativos tanto em redes sociais quanto em aplicativos de mensagens instantâneas, que se consolidaram como uma das ferramentas de comunicação mais importantes e populares [Karimpour et. al, 2021].

Esses ambientes digitais geram diariamente um volume expressivo de informações, refletindo mudanças significativas na forma como as pessoas consomem e compartilham notícias [Reis e Benevenuto, 2021]. Esse fenômeno é evidente em comunidades técnicas, onde a troca de experiências, soluções e referências ocorre em fluxo contínuo. O Telegram[1] combina atributos de mensageria e rede social (grupos temáticos públicos, histórico acessível), favorecendo comunidades de prática e a circulação de conhecimento aplicado em Ciência de Dados (Data Science – DS) e Inteligência Artificial (IA).

[1] https://web.telegram.org/

Apesar do papel crescente dessas comunidades, ainda são escassos estudos focados em grupos públicos de Telegram voltados a DS e IA no contexto brasileiro. Para entender que tipo de conhecimento é produzido e compartilhado nesses espaços, é necessário caracterizar os temas discutidos e como as perguntas se distribuem entre questões técnicas, recursos de estudo, formação acadêmica e inserção no mercado.

Diante desse contexto, este trabalho busca responder à seguinte questão de pesquisa: Quais categorias representam as principais discussões dos usuários em grupos públicos do Telegram focados em Ciência de Dados e Inteligência Artificial? Para respondê-la, desenvolvemos o Telegram4DS, um conjunto de 2.000 perguntas provenientes de 10 grupos públicos (amostragem aleatória), rotuladas manualmente via análise temática em quatro categorias: mercado, cursos, dúvidas gerais e material.

Contribuições: (1) dataset público e anonimizado (Telegram4DS) com 2.000 perguntas rotuladas de grupos brasileiros de DS/IA; (2) categorias temáticas das discussões e insights sobre padrões de demanda por suporte técnico, formação e mercado.

2. Trabalhos Relacionados

Diversos repositórios e estudos abordam conversas técnicas on-line, porém com escopos de plataforma e focos temáticos diferentes deste trabalho. O SOSum [Kou et. al, 2022] apresenta postagens do Stack Overflow com sentenças somativas rotuladas, buscando encorajar a construção de conjuntos de dados em maior escala com rótulos de alta qualidade para Stack Overflow e outros tipos. GitterCom [Parra et. al, 2020] oferece mensagens manualmente rotuladas e curadas de mensagens de desenvolvedores do Gitter voltadas a projetos open-source.

Em paralelo, Disco [Subash et. al. 2022] reúne conversas públicas do Discord em comunidades de desenvolvimento, com técnica de desembaraçamento (disentanglement technique) para extrair conversas das transcrições de chat. No âmbito do Telegram, Pushshift Telegram [Baumgartner et. al., 2020] agrega milhares de canais e centenas de milhões de mensagens, fomentando pesquisas em movimentos sociais e desinformação. O trabalho de Garimella e Tyson [2018] utiliza o WhatsApp com o objetivo de explorar a viabilidade de coletar esses dados para pesquisas em ciências sociais.

3. Metodologia

A coleta concentrou-se nos dados de grupos públicos do Telegram relacionados à discussão de DS e IA. Para a descoberta dos grupos relacionados ao tema deste trabalho, foi utilizada a metodologia definida em [Júnior et. al, 2021]: 1) Pesquisa de palavras-chave relevantes para encontrar grupos orientados ao tema; 2) Pesquisar URLs para grupos/canais do Telegram em outras plataformas sociais contendo palavras-chave selecionadas; 3) Ingresso nos grupos e canais relevantes com uma conta do Telegram; 4) Coleta de todas as mensagens, usuários e outros dados disponíveis dos grupos e canais.

A partir de um conjunto inicial de 28 grupos voltados a DS/IA no Telegram, selecionamos aleatoriamente 10 para compor o estudo, foram extraídas todas as mensagens dos grupos selecionados, desde o seu início até agosto de 2025. A data de criação dos grupos varia de 2016 a 2021.

As mensagens postadas nos grupos públicos do Telegram, são mantidas no histórico, o que significa que todos os usuários podem ver todas as mensagens, desde quando o grupo foi criado. Obtendo o histórico do grupo, o próximo passo para facilitar o processo de rotulagem foi criar um script em python[2] para converter as mensagens do formato JSON fornecido pelo histórico de mensagens para o formato CSV.

De um total de 631.014 mensagens, foram extraídas 200 perguntas por grupo (total: 2.000) por amostragem aleatória simples, evitando efeitos de sazonalidade ou eventos pontuais. Optou-se por rotular perguntas aleatórias em vez de mensagens consecutivas para garantir que o conjunto de dados representasse com maior precisão o comportamento dos grupos ao longo dos anos.

3.1. Descrição do Conjunto de Dados

O Telegram4DS está disponível on-line[3] no formato de arquivo csv. Cada linha contém um registro composto por seis campos de informações. As colunas são: 1) id grupo, 2) id mensagem, 3) data, 4) id usuário, 5) texto e 6) categoria. Foram consideradas questões de privacidade neste trabalho. Anonimizamos nomes de usuários e seus @s, para garantir a privacidade dos usuários e eliminar a possibilidade de identificação dos mesmos, também foram anonimizados nomes de instituições e locais. Foi atribuído a cada usuário um identificador único.

Tabela 1 Descrição do Conjunto de Dados.

Nome do Campo	Formato	Descrição
Id_Grupo	String	Identificador único do Grupo
Id_Mensagem	String	Identificador único da Mensagem
Data	Datetime	Data no formato ISO 8601
Id_Usuario	String	Identificador único do Usuário
Texto	String	Texto da Mensagem
Categoria	String	Categoria definida manualmente

Na primeira categoria, denominada material, observa-se uma busca predominante por fontes de conhecimento, como livros e artigos, que abordam determinados conteúdos específicos de interesse do usuário, onde encontrá-los e até mesmo o compartilhamento dos mesmos. A categoria material, onde normalmente as pessoas procuram ou são direcionadas para recursos como links, livros, artigos ou conjunto de dados e as mensagens podem ser consideradas um repositório relevante para usuários do nível básico ao avançado terem acesso à fonte mais abrangente de conhecimento [Karbasian e Johri, 2020]. A Tabela 2 ilustra exemplos dessa categoria.

Tabela 2. Exemplos da Categoria Material.

MATERIAL

[2] https://www.python.org/

[3] https://github.com/garganoleo/TELEGRAM4DS

Boa noite pessoal! Alguém tem o livro introdução a estatística do Triola que possa compartilhar???
Olá, como vao? Conhecem algum material sobre a API Cloud vision para recomendar? A rede neural Yolo e nem outra plataforma não é uma opçao para o momento. Obrigado!
Alguém tem indicação de livros de engenharia de dados?

A segunda categoria, denominada cursos, agrupa perguntas sobre cursos de nível superior, pós-graduação e bootcamps, visto que são pedidas sugestões para diversos níveis de conhecimento e sobre várias tecnologias e competências, além de preocupação com o valor, observa-se que alguns usuários buscam opções de menor custo, enquanto outros utilizam o grupo para avaliar custo-benefício para se certificar de que vale a pena o investimento. O interesse em questões relacionadas a educação por usuários fora da área de Ciência da Computação pode ser visto como no trabalho de [Tacheva et. al, 2022] através das lentes da teoria econômica, que prevê que em tempos de turbulência econômica, há um aumento no número de pessoas que buscam melhorar sua compatibilidade com as condições de mercado, aprofundando sua educação ou fazendo a transição para novos campos. Em relação ao questionamento sobre as Instituições, a disseminação de informações pode ser favorável às empresas, mas também pode prejudicar suas reputações [Lueg, 2007]. A Tabela 3 ilustra exemplos dessa categoria.

Tabela 3. Exemplos da Categoria Cursos.

CURSOS
Boa dia, pessoal. Alguém conhece ou já fez a mentoria do XXXX?
Alguem no grupo faz pós ou MBA em data science? Pode compartilhar a experiencia?
Galera, acham que o conteúdo da XXXX é bom para se especializar na área de Dados? Ou valeria mais juntar uma grana e comprar um curso mais caro?

A terceira categoria, denominada mercado, os usuários demonstram dúvidas relacionadas ao salário na área, formas de atrair o interesse de recrutadores de empresas, migração de áreas, vagas de emprego e diversos questionamentos sobre a situação atual do mercado de trabalho. A predominância da categoria mercado pode estar associada à procura por adaptação às mudanças no estilo de vida e à transição para o uso cada vez maior de dados em áreas não relacionadas a Ciência da Computação, a procura por migração de carreira em resposta às mudanças econômicas, ao impacto negativo tanto nos profissionais da indústria quanto na academia com relação à segurança no emprego bem como às perspectivas de carreira associados à busca por trabalhos remotos/híbridos. Isso indica que mensagens instantâneas são importantes para que desenvolvedores se mantenham informados sobre eventos e novas oportunidades de emprego [Silva, 2022]. A Tabela 4 ilustra exemplos dessa categoria.

Tabela 4. Exemplos da Categoria Mercado.

MERCADO
Pessoal, os que já estão no mercado de trabalho, qual deve ser a expectativa salarial pra um primeiro trabalho?
Sou da área da logística mas estou pensando em fazer curso de analista de dados. Não sei nada da área, mas tenho interesse em aprender e levar como profissão. Gostaria de saber como está a demanda para essa área.
Bom dia senhores, alguma oportunidade de DS ou BI Jr/Trainee ??

A quarta categoria, denominada dúvidas gerais, percebe-se dúvidas relacionadas a formas de solucionar necessidades pontuais, sejam tecnológicas ou matemáticas, além de questões sobre desempenho, tecnologias e métodos. É a categoria mais "técnica" do nosso conjunto de dados. Essa é a categoria mais abrangente, englobando diferentes padrões de perguntas. A Tabela 5 ilustra exemplos dessa categoria.

Tabela 5. Exemplos da Categoria Dúvidas Gerais.

DÚVIDAS GERAIS
Vocês montam a pipeline em um script ou com orquestrador ?
Pessoal, qual a melhor forma pra lidar com dados em escalas muito diferentes? Só Normalizar já ajuda? OU Vale jogar em log tbm?
Bom dia pessoal, alguém usando airflow e spark no k8s??

A Tabela 6 exibe o número de perguntas por categorias no Telegram4DS, observa-se que as dúvidas gerais representam quase 55% do conjunto de dados e que as mensagens relacionadas a cursos e material representam mais de 30% dos dados. Em concordância com os autores [Storey et. al. 2014], nossa pesquisa ilustrou que os dados do Telegram4DS demonstram que a busca por sites de aprendizado online promove mecanismos adicionais de aprendizagem em comunidades online.

Tabela 6. Categorias do Telegram4DS.

Categoria	Quantidade	%
Cursos	380	19
Dúvidas Gerais	1076	53,8
Material	284	14,2
Mercado	260	13

4. Limitações

Reconhecemos que as nossas categorias e assuntos são um retrato das mensagens dos grupos de Telegram, influenciados pela nossa compreensão e interpretação dos dados. Uma vez que novos grupos podem aparecer e grupos existentes podem desaparecer, por mais que outros investigadores possam interagir com os dados e identificar temas de forma diferente, reconhecemos que as replicações do nosso estudo podem encontrar temas diferentes. Para trabalhos futuros, sugerimos que mais pesquisadores analisem as categorias.

Constatamos que os resultados do nosso trabalho podem ter sido influenciados pela seleção de grupos realizada no estudo. O Telegram não fornece uma lista de grupos, nesse trabalho não selecionamos todos os grupos que tratam sobre o tema Ciência de Dados e Inteligência Artificial da plataforma. Os resultados do nosso estudo podem não ser generalizáveis para outras plataformas de chat ou comunicações de desenvolvedores. Uma possível expansão para conjuntos de dados maiores também pode levar a resultados de avaliação diferentes.

Também é possível que o conjunto de dados para análise manual não seja representativo do corpus completo de conversas do Telegram para uma determinada comunidade. O tamanho dos conjuntos de dados foi escolhido para fornecer uma amostra estatisticamente representativa, viável para análise do autor. No entanto, a

escala para conjuntos de dados maiores para a análise manual pode levar a resultados diferentes.

5. Conclusões e Trabalhos Futuros

Este trabalho apresenta o Telegram4DS, um conjunto de 2.000 perguntas rotuladas de 10 grupos públicos do Telegram focados em Ciência de Dados e Inteligência Artificial no Brasil. A partir de 631.014 mensagens, foi realizada a anonimização e análise temática, organizando as perguntas nas categorias material, cursos, mercado e dúvidas gerais. Como trabalhos futuros, pretende-se expandir o dataset para mais perguntas e incluir novos grupos; incluir camadas adicionais de anotação (intenção, dificuldade, área técnica); treinar e disponibilizar classificadores automáticos e sumarizadores; comparar Telegram vs. Discord/WhatsApp.

Referências

Baumgartner, J., Zannettou, S., Squire, M. and Blackburn, J. 2020. The Pushshift Telegram Dataset. Proceedings of the International AAAI Conference on Web and Social Media. 14, 1 (May 2020), 840-847. DOI:https://doi.org/10.1609/icwsm.v14i1.7348

Karimpour, D.; Zare Chahooki, M. A. and Hashemi, A. "User recommendation based on Hybrid filtering in Telegram messenger," 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 2021, pp. 1-7, doi: 10.1109/CSICC52343.2021.9420562.

Parra, E. et. al. 2020. GitterCom: A Dataset of Open Source Developer Communications in Gitter. In Proceedings of the 17th International Conference on Mining Software Repositories (MSR '20). Association for Computing Machinery, New York, NY, USA, 563–567. https://doi.org/10.1145/3379597.3387494

Garimella, K., and Tyson, G. 2018. Whatapp Doc? A First Look at Whatsapp Public Group Data. In Twelfth Interna tional AAAI Conference on Web and Social Media.

Tacheva, J.; Lahiri, S. and Saltz, J. "Analyzing a Data Science Online Practitioner Community: Trends and Implications for Data Science Project Management," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 2673-2681, doi: 10.1109/BigData55660.2022.10020600.

Reis, J. C. S. and Benevenuto, F. 2021. Supervised Learning for Misinformation Detection in WhatsApp. In Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia '21). Association for Computing Machinery, New York, NY, USA, 245–252. https://doi.org/10.1145/3470482.3479641

Karbasian, H., & Johri, A. (2020, February). Insights for curriculum development: Identifying emerging data science topics through analysis of Q&A communities. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education (pp. 192-198).

Kou, Y.; Gray, C. M.; Toombs, A. L. and Adams, R. S. "Understanding social roles in an online community of volatile practice: A study of user experience practitioners on reddit", ACM Transactions on Social Computing, vol. 1, no. 4, pp. 1-22, 2018.

Lueg, C.P. (2007), Querying information systems or interacting with intermediaries? Towards understanding the informational capacity of online communities. Proc. Am. Soc. Info. Sci. Tech., 44: 1-6. https://doi.org/10.1002/meet.1450440249

Júnior, M.; Melo, P.; Silva, A. P. C.; Benevenuto, F. and Almeida, J. 2021. Towards Understanding the Use of Telegram by Political Groups in Brazil. In Proceedings of the

Brazilian Symposium on Multimedia and the Web (WebMedia '21). Association for Computing Machinery, New York, NY, USA, 237–244. https://doi.org/10.1145/3470482.3479640

Silva, C. M. C. Identifying reusable knowledge in developer instant messaging communication. 2022

Subash, K. M., Kumar, L. P., Vadlamani, S. L., Chatterjee, P., & Baysal, O. (2022, May). DISCO: A dataset of Discord chat conversations for software engineering research. In Proceedings of the 19th International Conference on Mining Software Repositories (pp. 227-231).