|  |  |  |
| --- | --- | --- |
| **Template do Dataset card** | | |
|  |  |  |
| **Campo** | **Descrição do Campo** | **Dataset** |
| **1. Detalhes** | | |
| Nome |  | Paul/hatecheck-portuguese |
| Descrição | Deve conter .... (completar) | É um conjunto de testes funcionais para modelos de detecção de discurso de ódio. |
| Proprietário | Criador ou Responsável pelo DataSet | Paul Röttger |
| Versão | Identificador da versão | v1 |
| Data de criação |  | 20/jun/22 |
| Data da última atualização |  | 20/jun/22 |
| URL |  | <https://huggingface.co/datasets/Paul/hatecheck-portuguese> |
| Requisitos de Acesso | Requisitos de acesso, contato do responsável (texto) | paul@rewire.online; paul.rottger@unibocconi.it |
| Licença de uso | GPL, creative commons etc. | cc-by-4.0 |
| **2.Características** | | |
| Tipo | Lista: Estruturado, Texto, Imagem, Texto-Imagem, Outros | Estruturado |
| Domínio | Descrever a área de domínio do corpus | Tweets |
| Volume | Tamanho total (em bytes) | 997 KB |
| Documentos | Descrição da composição. Lista: TXT, PDF, IMAGEM, json, outros | TXT |
| Dados quantitativos | Descrição quantitativa e qualitativa da composição. Lista: quantidade de documentos, sentenças, tokens, ... | 3,691 linhas. O comprimento médio do caso de teste é de 8,5 palavras (padrão = 3,6) ou 46,0 caracteres (padrão = 21,3). |
| Sanidade | Descrição de dados faltantes, com ruído, incompletos, etc (texto) | Tem dados faltantes |
| Método de extração | Descrição do processo (roteiro,scripts, etc) | Coletaram 42.930 tweets com base em uma pesquisa de 29 perfis de usuários, 19 palavras-chave e dez hashtags, no ano de 2017. Eles então filtraram os tweets, mantendo apenas os tweets em português e removendo duplicatas e retuítes, resultando em 33.890 tweets. Por fim, eles estabeleceram um limite máximo de 200 tweets por instância de pesquisa para manter uma fonte mais diversificada de tweets. O conjunto de dados final de 5.668 tweets. |
| **3.Anotações** | | |
| Tarefa | Informa a tarefa para a qual a anotação foi realizada. Classificação, Reconhecimento de Entidades Nomeadas (NER), ... | Classificação de Texto |
| Classes | Descreve as classes anotadas. Deve incluir o label das classes, o significado de cada label e a distribuição delas no dataset. (texto) | odioso, não odioso (hateful, non-hateful) |
| Balanceamento | Classes distribuídas de forma igualitária em termos de amostras (binário: SIM/NÃO) | Não |
| Tipo | Deve indicar a forma como o dado foi anotado: manual, automático ou semiautomático. Manual: quando a anotação foi realizada por anotadores humanos; Automático: quando a anotação foi realizada por um algoritmo; Semiautomático: quando a anotação foi realizada por um algoritmo e, posteriormente, revisada manualmente.  (texto) | Manual |
| Método | Quando for do tipo Manual deve mencionar a quantidade de anotadores e as suas expertises. Deve descrever ou ter um link para o GuideLine utilizado no processo de anotação manual. Quando for do tipo automático deve ter preferencialmente o programa (script) que realizou a anotação automática, um link para este classificador ou nome da ferramenta de anotação. E, por fim, quando for semiautomático, deve ter tanto informações sobre o classificador (programa anotador) quanto dos anotadores humanos e os guias usados no processo de revisão. (texto) | 18 anotadores |
| Qualidade | Indica a qualidade da anotação. No caso de anotação manual, deve informar a métrica usada para medir a concordância dos anotadores (Ex: Kappa). No caso de anotadores automáticos, deve informar as métricas do classificador, tais como Acurácia, Precisão, Recall e F-Measure. (texto) | Todos os tweets no conjunto de dados foram anotados como odiosos ou não odiosos por 18 falantes nativos de português não especialistas foram contratados. Cada tweet foi anotado por três anotadores, e a concordância entre os anotadores foi baixa, com um Kappa de Cohen de 0,17. |
| Idioma | Informar o idioma. Valor default desse campo: Português. (texto) | Português |
| Formato | Descreve o formato usado nos arquivos do dataset: CSV, IOB, IOB2, XML, JSON,... Recomendável seguir o formato pré-definido para o Sinapses. Em caso contrário, é recomendável incluir um exemplo do formato utilizado. (texto) | CSV |
| Outras informações | Mencionar aqui outras informações relevantes, tais como técnicas usadas para melhorar o desbalanceamento (Ex: Data Augmentation). (texto) | 200 twits por instancia. |
| Material de Referência | Indicar documentos nos quais pode-se obter ainda mais informações, tais como: publicações, sites, ... (texto) | https://arxiv.org/abs/2206.09917 ; https://aclanthology.org/W19-3510/ |
| Local | Para o caso dele não estar armazenado na plataforma Sinapses. Neste caso, indicar o link aqui. (texto) | - |
| **4.Observações** | | |
| Aspectos éticos/legais | Descrição se os dados observam as regras da LGPD, contém anonimização e possuem tratamentos de viés de natureza social (gênero, classe sócio-econômica, etc) ou técnica (áreas ou sub-áreas jurídicas) | Levaram em consideração a privacidade dos autores das mensagens coletadas anonimizado, omitindo o ID do tweet. Como consequência, é possível alcançar o tweet original e o usuário apenas pesquisando o texto exato do tweet. Disponibilizaram o conjunto de dados no GitHub apenas para fins de pesquisa, sob a condição de que essa pesquisa não seja realizada. |
| Outras observações | Descrever aqui qualquer outra informação que for relevante. | - |