

Da Definição das Entidades à Anotação do Corpus: Desafios da tarefa no Direito do Consumidor

Samuel Rios da Silva¹, Aline Athaydes¹, Babacar Mane¹, Daniela Barreiro Claro¹

¹FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos – Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{samuelrs, alineathaydes, mbabacar, dclaro}@ufba.br

Abstract. *Data annotation is a crucial step in the development of machine learning models aimed at Named Entity Recognition (NER). In this context, given the identified need and the lack of specialized NER models for the Brazilian Consumer Defense Code (CDC), this work presents the process of annotating a corpus with specific entities related to the CDC, as well as the challenges encountered, the lessons learned, the main results obtained, and our next steps.*

Resumo. *A anotação de dados é uma etapa crucial para a construção de modelos de aprendizado de máquina voltados ao Reconhecimento de Entidades Nomeadas (REN). Nesse sentido, percebida a necessidade e a ausência de Modelos REN especialistas no contexto do Código de Defesa do Consumidor (CDC), este trabalho apresenta o processo de anotação de um corpus com entidades específicas relacionadas ao CDC, assim como os desafios encontrados, os aprendizados, principais resultados obtidos e nossos próximos passos.*

1. Introdução

Os modelos de Reconhecimento de Entidades Nomeadas (REN) são capazes de identificar e classificar entidades mencionadas em um grande volume de dados textuais, facilitando sua análise semântica. Nesse contexto, a anotação de um *corpus* para o treinamento desses modelos é uma atividade essencial de interpretação e classificação humana e/ou automática, que leva em consideração o contexto linguístico [Caseli and Nunes 2026].

No âmbito do Código de Defesa do Consumidor (CDC) [Brasil 1990], um tópico específico do domínio jurídico, a literatura lista um conjunto de corpora voltados para o REN jurídico [Caseli and Nunes 2026]. Em sua maioria, são compostos por documentos formais e possuem entidades genéricas ou específicas de outros tópicos, não abrangendo entidades representativas adequadas para o CDC.

Sendo o CDC uma referência central para os direitos do consumidor [Athaydes et al. 2024, Athaydes et al. 2025], identificamos uma lacuna na geração de um conjunto de etiquetas especializadas nesse código, focado na entrada de textos redigidos em linguagem coloquial por usuários. Este trabalho, portanto, apresenta o processo de criação dessas etiquetas e anotação de um *corpus* com entidades relacionadas ao CDC, destacando os desafios encontrados, os resultados obtidos e as lições aprendidas ao longo do processo.

Este artigo está estruturado como segue: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 descreve a construção do *corpus*. As Seções 4 e 5 apresentam o

conjunto de entidades e a ferramenta de anotação respectivamente. A Seção 6 descreve os resultados e a conclusão é apresentada na Seção 7.

2. Trabalhos Relacionados

Nos últimos anos, trabalhos avançaram sobre documentos legislativos brasileiros. O UlyssesNER-Br ampliou os recursos para PLN jurídico ao explorar textos legislativos como fonte para anotação de entidades [Albuquerque et al. 2022]. Nunes et al. [Nunes et al. 2024] também investigaram REN em textos legislativos em português com autoaprendizado. Esses estudos indicam que o REN jurídico em português tem se concentrado principalmente em documentos formais e institucionais.

Apesar desses avanços, ainda há uma lacuna quando o foco se desloca de documentos formais para relatos de consumidores escritos em linguagem coloquial. No contexto do Código de Defesa do Consumidor (CDC), entidades como consumidor, fornecedor, produto, serviço e relação de consumo podem aparecer de forma implícita, incompleta ou ambígua, exigindo critérios de anotação específicos. Assim, este trabalho propõe uma definição de etiquetas especializadas e apresenta os desafios de anotação dos relatos dos consumidores, aplicando a tarefa de REN jurídico a textos produzidos por usuários finais no domínio do CDC.

3. Construção do Corpus

Textos jurídicos, assim como textos do CDC, contêm frequentemente linguagem técnica e sentenças longas [Bessa et al. 2026], dificultando a compreensão por grande parte de seus leitores. Com o intuito de obter um conjunto de dados mais próximos aos descritos pelos usuários, optou-se pela plataforma *Consumidor.gov.br*, que reúne relatos (ou reclamações) redigidos na linguagem do usuário final. O *corpus* foi obtido por meio da extração de mil relatos de consumidores com as suas respectivas relações de consumo.

4. Conjunto de Entidades

As entidades a serem anotadas foram definidas e extraídas diretamente dos artigos 2º e 3º do CDC [Brasil 1990]:

- **Consumidor:** Toda pessoa física ou jurídica que adquire ou utiliza produto ou serviço como destinatário final.
- **Fornecedor:** Toda pessoa física ou jurídica, pública ou privada, nacional ou estrangeira, bem como os entes despersonalizados, que desenvolvem atividade de produção, montagem, criação, construção, transformação, importação, exportação, distribuição ou comercialização de produtos ou prestação de serviços.
- **Produto:** Qualquer bem, móvel ou imóvel, material ou imaterial.
- **Serviço:** Qualquer atividade fornecida no mercado de consumo, mediante remuneração, inclusive as de natureza bancária, financeira, de crédito e securitária, salvo as decorrentes das relações de caráter trabalhista.

Com essas definições iniciais estabelecidas, foi realizada uma primeira reunião com uma especialista no domínio jurídico para validação desse conjunto de entidades. Embora as categorias Consumidor, Fornecedor, Produto e Serviço estejam definidas nos artigos 2º e 3º do CDC, observou-se que, nos relatos analisados, a identificação do consumidor muitas vezes é implícita e está associado à relação estabelecida entre o usuário

e o fornecedor. Como resultado, observou-se a necessidade de identificar os eventos que estabelecem a relação de consumo. Foi, então, adicionada o evento *RelacaoConsumo*, definida como segue:

- **RelacaoConsumo**: Um evento que estabelece uma relação de consumo, por exemplo: "comprei" ou "contratei".

O acréscimo de tal evento, estende a tarefa de REN para incluir o reconhecimento de Eventos [Caseli and Nunes 2024].

5. Ferramenta de Anotação

Um total de 50 textos aleatórios foram selecionados para anotação manual e validação por duas especialistas. As anotações ocorreram através da ferramenta Label Studio [Tkachenko et al. 2025], uma plataforma de código aberto que permite selecionar trechos do texto e associá-los às entidades. A Figura 1 exibe sua interface já com o conjunto de etiquetas definidos para essa tarefa.

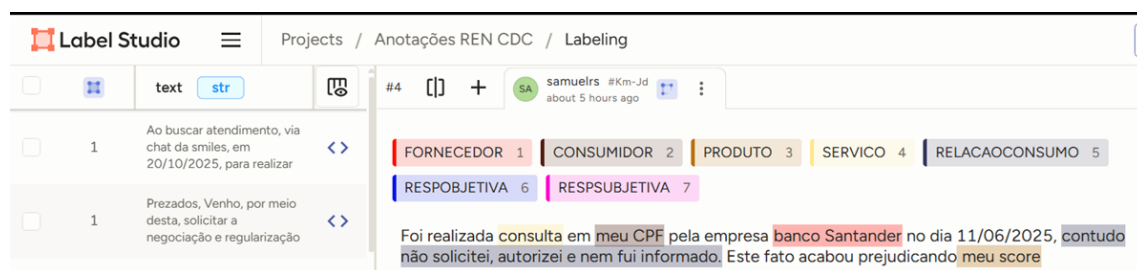


Figura 1. Interface do Label Studio para anotação dos relatos do Consumidor.gov.br

6. Resultados e Discussões

Para avaliar a concordância entre as anotadoras, adotaram-se duas métricas: o Kappa de Cohen (κ) adaptado para janelas de sobreposição (spans) e o F1-Score Cruzado (Pairwise F1). O resultado para o Kappa de Cohen foi de $\kappa = 0,1074$ (10%) e para o F1-Score Cruzado obtivemos um resultado de 0,3088 (30%). Tais resultados apontam para uma baixa concordância geral entre as anotadoras.

Após a anotação, uma discussão com as especialistas buscou mapear os motivos dessa divergência que seguem descritos. O primeiro problema observado foi a falta de um **manual de anotação (guideline)**, fazendo com que as especialistas concluíssem que era obrigatória a anotação das entidades *Consumidor*, *Fornecedor* e *Produto*. Esse entendimento equivocado por parte das especialistas favoreceu com que cada uma adotasse **convenções próprias** para identificar tais entidades que não estavam presente explicitamente no relato (por exemplo, “meu CPF” era marcado como *Consumidor*). Além disso, houve **interpretações divergentes** sobre a entidade *RelacaoConsumo*, e **trocas frequentes entre os conceitos** de *Produto* e *Serviço* entre ambas as especialistas.

A partir dessa etapa, observou-se a necessidade de um alinhamento prévio em conjunto com ambas as especialistas. Uma vez alinhado em um segundo momento e as dúvidas sanadas, criou-se um *Manual de Anotação* formalizando as regras que deveriam

ser seguidas de acordo com o CDC. A última versão desse manual, abrange as seguintes entidades: *Consumidor*, *Fornecedor*, *Produto*, *Serviço*, *RelaçãoConsumo*. O manual está em fase final de refinamento para aprimorar essas definições e solucionar discussões em aberto.

Tabela 1. Resultado das Métricas por Entidade

Entidade	Kappa de Cohen	F1-Score Cruzado	Instâncias (Anotadora 1)	Instâncias (Anotadora 2)
CONSUMIDOR	0,4728	0,5591	44	49
FORNECEDOR	0,7764	0,8095	41	43
PRODUTO	0,0368	0,0435	1	45
RELACAOCONSUMO	-0,1140	0,0808	49	50
SERVICO	0,1132	0,2581	49	50

A partir do manual e o alinhamento entre as especialistas, as concordâncias entre as anotadoras foram estratificadas por etiquetas (Tabela 1), e observou-se um aumento significativo nas entidades *Fornecedor* e *Consumidor*. A entidade *Serviço* ainda carece de análise assim como a entidade *Produto*. A pior discordância foi observada na *RelaçãoConsumo*, visto que não há um acordo entre as especialistas, sendo necessária uma nova rodada de alinhamento ou a retirada dessa etiqueta do processo de anotação. Assim, o presente trabalho apresentou os desafio de uma tarefa de anotação, mesmo quando se tem a presença de especialistas no domínio.

7. Conclusões e Trabalhos Futuros

A análise preliminar revelou baixa concordância inicial, mas o F1-Score de 80% para *Fornecedor* demonstra o potencial da tarefa. Ficou claro que o maior desafio ao transpor conceitos do CDC para a linguagem coloquial é definir um conjunto representativo de entidades com precisão.

Essa baixa concordância derivou da falta de um manual e de alinhamento prévio, gerando interpretações subjetivas e confusões (como entre *Produto* e *Serviço*). Isso evidencia a necessidade de ciclos iterativos; o alinhamento posterior com especialistas para criar o Manual de Anotação foi essencial para padronizar o processo.

Como trabalhos futuros, prevê-se um aumento significativo na concordância entre anotadoras para consolidar um *corpus* ouro (*gold standard*). A partir dele, planeja-se a anotação automática dos dados restantes via Grandes Modelos de Linguagem (LLMs), com revisão humana para garantir confiabilidade e agilizar a construção do *corpus* especialista no CDC.

Agradecimentos

Os autores agradecem o financiamento parcial através da bolsa FAPESB 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, FAPESB CCE 0022/2023, FAPESB TIC 002/2015. Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). Gostaríamos de agradecer ao Escavador pela parceria neste trabalho.

Referências

- Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitória, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. a. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 3–14, Berlin, Heidelberg. Springer-Verlag.
- Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D. B., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (QA) models. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 522–529, Belém do Pará, Brazil. Association for Computational Linguistics.
- Athaydes, A., Mota, L., Neto, F. M., Silva, S., Mane, B., Claro, D., Souza, M., and Lisboa, A. (2025). Towards a corpus methodology for llms in the legal domain. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 272–282, Porto Alegre, RS, Brasil. SBC.
- Bessa, A. C. C., Lobato, F. M. F., and Junior, A. F. L. J. (2026). Development and evaluation of a hybrid information retrieval system applied to the Brazilian legal domain. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 186–190, Salvador, Brazil. Association for Computational Linguistics.
- Brasil (1990). Lei nº 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial da União, Brasília, DF.
- Caseli, H. d. M. and Nunes, M. d. G. V. (2024). Extração de informação. In Caseli, H. d. M. and Nunes, M. d. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações*, chapter Extração de Informação. 3º edition.
- Caseli, H. M. and Nunes, M. G. V., editors (2026). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3. BPLN, 4 edition.
- Nunes, R. O., Balreira, D. G., Spritzer, A. S., and Freitas, C. M. D. S. (2024). A named entity recognition approach for Portuguese legislative texts using self-learning. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 290–300, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Tkachenko, M., Malyuk, M., Holmanyuk, A., and Liubimov, N. (2020-2025). Label Studio: Data labeling software. Open source software available from <https://github.com/HumanSignal/label-studio>. Acesso em: 02 jun. 2026.