PromptNER: Uma Abordagem para Reconhecimento de Entidades Nomeadas em Dados Sensíveis a Partir de Instâncias Rotuladas Automaticamente

Claudio M. V. de Andrade[1], Celso França[1], Fabiano Belém[1], Gabriel Jallais[1], Marcelo A. S. Ganem[1], Gabriel Teixeira[1], Alberto H. F. Laender[1], Marcos A. Gonçalves[1]

[1]Unversidade Federal de Minas Gerais (UFMG)

{claudio.valiense,celsofranca,fmuniz,gabrieljallais}@dcc.ufmg.br

{marceloganem,gabrielmedeiros,laender,mgoncalv}@dcc.ufmg.br

Abstract. *In this article, we address the task of Named Entity Recognition (NER) for Organizations and Products/Services in textual complaints recorded on web platforms. Due to the high inference power of Large Language Models (LLM's), there is a growing interest in their application. However, they face issues of high infrastructure cost and privacy concerns when using external API's. Thus, we propose an approach that uses LLM's for the recognition of entities in complaints and then trains simpler models, such as the SpERT method. The enhanced NER model achieves significant gains of 41% to 129% in F-score compared to the labeled data-only model.*

Resumo. *Neste artigo, abordamos a tarefa de Reconhecimento de Entidades Nomeadas (REN) nos casos de Organizações e Produtos/Serviços presentes em reclamações textuais registradas em plataformas na Web. Devido ao alto poder de inferência dos modelos de linguagem de larga escala (LLM's), há interesse crescente em sua aplicação, porém eles enfrentam problemas de alto custo de infraestrutura e privacidade ao utilizar API's externas. Assim, propomos uma abordagem que utiliza LLM's para o reconhecimento de entidades nas reclamações e que, em seguida, treina modelos mais simples, como o método SpERT. O modelo de REN aprimorado obtém ganhos significativos de 41% a 129% em F-score em comparação com o modelo de dados rotulados apenas manualmente.*

1. Introdução

Diariamente, milhares de relatos de consumidores são registrados em plataformas como Consumidor.gov.br[1] e Procon[2]. Reconhecer nesses relatos entidades do tipo **Organização** e **Produto** ou **Serviço** é essencial para que os órgãos de controle tomem as medidas cabíveis para proteger os direitos dos consumidores. Entretanto, devido ao grande volume de dados registrados nessas plataformas, é impossível que essa extração seja feita manualmente, o que motiva o estudo de um problema de processamento de linguagem natural (PLN) conhecido como Reconhecimento de Entidades Nomeadas (REN).

[1]https://www.consumidor.gov.br
[2]https://www.procon.pr.gov.br

Neste artigo, abordamos um cenário real bem desafiador do Ministério Público de Minas Gerais (MPMG), no qual é preciso reconhecer entidades nomeadas classificadas como sendo dos tipos Organização e Produto ou Serviço, a partir de um conjunto de manifestações de consumidores sem qualquer rótulo. Uma vez que tais entidades agregam informações importantes contidas em textos, REN é uma tarefa útil em diversas aplicações tais como deduplicação de registros [Silva et al. 2019, Mangaravite et al. 2022], integração de dados [Brunner & Stockinger 2020], construção de bases de conhecimento [Niu et al. 2012] e busca em coleções de textos não-estruturados [Caputo et al. 2009].

Abordagens baseadas em *transformers* como, por exemplo, *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2019], constituem o estado-da-arte em termos de métodos para REN. Dentre essas abordagens, destaca-se o modelo discriminativo SpERT (*Span-based Entity and Relation Transformer)* que é adotado neste artigo. O SpERT baseia-se em aspectos semânticos (i.e., contextuais) de sequências de palavras fazendo com que alcancem alta eficácia, em vários conjuntos de dados de referência, com um custo computacional relativamente baixo.

Entrentanto, um dos maiores desafios dos métodos discriminativos é a dependência de uma quantidade relativamente grande de dados rotulados, algo que inexiste para o conjunto de dados disponibilizado pelo MPMG. A obtenção de dados rotulados é geralmente feita manualmente, apresentando um alto custo financeiro em termos de recursos humanos (horas trabalhadas), além de possibilitar a inserção de ruídos. Neste artigo, tratamos esse desafio por meio de uma abordagem de aprendizagem de máquina baseada em *prompt-learning* [Luo et al. 2022, Ye et al. 2023]. Tal abordagem é relativamente nova, particularmente nos casos em que modelos de linguagem de larga escala (LLM's, da sigla em inglês para *Large Language Models*) são explorados (ex., ChatGPT[3], BLOOM[4]). Mais especificamente, este tipo de abordagem utiliza-se de um *prompt* (instrução, frase ou texto inicial) que é apresentado ao modelo de linguagem para iniciar uma conversa ou realizar uma tarefa.

No contexto da tarefa REN, o *prompt* pode ser usado para identificar entidades nos textos das manifestações. Por exemplo, ao empregar um LLM para identificar nomes de organizações, podemos fornecer alguns exemplos de manifestações rotuladas (nas quais as entidades estão identificadas), seguidos por uma manifestação não rotulada. O LLM então infere a partir dos exemplos as entidades na manifestação não rotulada. A Figura 1 apresenta uma requisição no formato de um *prompt* submetida ao BLOOM. Nessa requisição não são fornecidas as palavras "mercado livre" em vermelho, terminando a requisição em "[Organização]:". Entretanto, como resposta, o modelo generativo retorna o texto completado com o termo "mercado livre".

Uma vez aumentada a quantidade de relatos rotulados, o aprendizado baseado em *prompt* pode ser combinado com outras técnicas, como aprendizado por transferência [de Andrade et al. 2023] ou ativo [Silva et al. 2022]. Assim, propomos uma abordagem que combina *prompt-learning* com *fine-tuning* em duas etapas: (i) uma rotulação inicial de entidades nomeadas é realizada usando a técnica de *prompt-learning* em LLM's e (ii) os dados rotulados assim gerados são utilizados para ajustar (*fine-tuning*)

[3] `https://chat.openai.com`

[4] `https://huggingface.co/bigscience/bloom`

[Texto]: Estou recebendo cobranças da Claro de produtos que não reconheço. Vide tela em anexo. CPF: xxx.xxx.xxx-xx
[Organização]: Claro
[Texto]: Estou a pouco tempo no mercado livre , minha primeira venda e já estou com problema . o mercado livre suspendeu a minha conta , sem motivos ! preciso que resolvam meu problema o mais rápido possível. indignada com a plataforma !
[Organização]: mercado livre

Figura 1. Exemplo de um *prompt* fornecido e gerado pelo BLOOM. O texto em preto foi o fornecido ao BLOOM, enquanto que o texto em vermelho foi gerado por ele.

modelos de linguagem mais simples, econômicos e "preservadores de privacidade", visto que é viável manter esses modelos na própria infra-estrutura da aplicação, mesmo sendo ela de pequeno ou médio porte, como é o caso do SpERT para a tarefa e o domínio específico considerados (REN em relatos de consumidores). Esta última característica é da maior importância quando consideramos o contexto dos orgãos e empresas públicas, no qual ambos estão sujeitos à Lei Geral de Proteção de Dados Pessoais (LGPD).

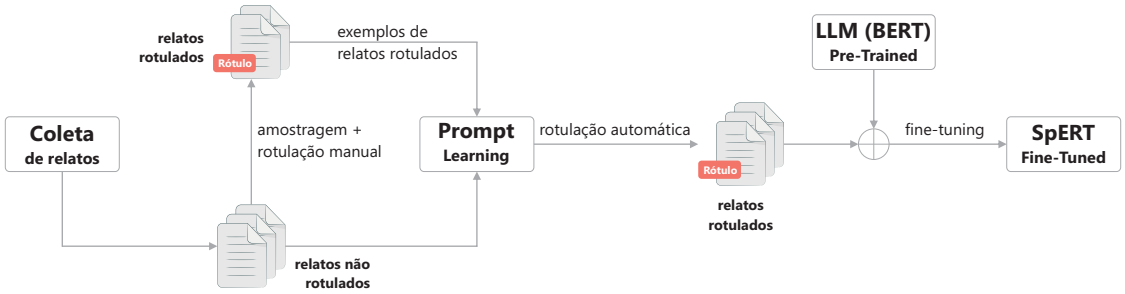

Figura 2. Abordagem proposta.

A abordagem aqui proposta envolve três etapas, conforme apresentado na Figura 2. A primeira etapa, "Coleta de Relatos", visa coletar as reclamações referentes à utilização da plataforma **Consumidor.gov.br**, cujos textos não incluem qualquer marcação das entidades reclamadas. A segunda etapa, denominada "Prompt-Learning", utiliza um modelo generativo para identificar as entidades mencionadas na reclamação. Finalmente, a terceira e última etapa executa um processo de *Fine-Tuning* que visa ajustar um modelo que é o estado da arte na tarefa REN para o domínio especifico de reclamações referentes a organizações e a produtos ou serviços. Ao final dessas três etapas temos um modelo capaz de rotular as reclamações sem a utilização de modelos generativos, evitando assim custos adicionais ou problemas com a proteção de dados sensíveis. Essas etapas são explicadas com mais detalhes na Seção 4.

Em resumo, neste artigo pretendemos responder às seguintes questões de pesquisa:

QP1: Qual o nível de concordância entre os avaliadores sobre a rotulação de relatos de consumidores usando um modelo generativo?
QP2: Qual é a eficácia da rótulação obtida a partir de LLM's como o BLOOM?
QP3: Qual é a eficácia dos modelos do estado-da-arte como o SpERT em dados rotulados a partir de modelos generativos?

Em síntese, a QP1 é respondidada através da proporção de concordância e por meio do coeficiente alfa de Krippendorff, o qual é usado para medir a concordância entre

os avaliadores. Para o caso da QP2, calculamos a precisão, a revocação e a medida-f (*f-score*) de cada classe, além de apresentar exemplos de casos de acerto e de erro. Para responder à última questão de pesquisa (QP3), calculamos também a precisão, a revocação e a medida-f, para então realizamos uma comparação entre o cenário onde é possível que apenas poucos dados sejam rotulados manualmente (Cenário-1) e o cenário onde temos dados rotulados a partir de uma estratégia de *prompt-learning* (Cenário-2). Todas essas questões serão respondidas com mais detalhes nas próximas seções.

Em suma, as contribuições deste artigo são:

1. Proposta e avaliação de uma etapa de rotulação automática de entidades nomeadas baseada em LLM's;
2. Melhoria de modelos para REN mais econômicos, escaláveis e "preservadores de privacidade" de dados sensiveis, como o SpERT, por meio de um processo de *fine-tuning* com os dados produzidos na etapa de rotulação;
3. Aplicação e avaliação das etapas descritas nos itens (1) e (2) no reconhecimento de organizações, produtos e serviços em relatos de consumidores, com resultados que mostram os benefícios dos métodos propostos.

2. Trabalhos Relacionados

Esta seção apresenta uma visão geral dos trabalhos relacionados ao problema de reconhecimento de entidades nomeados com foco em dois tipos específicos, abordagens discriminativas e abordagens generativas. A diferença principal entre esses dois tipos de abordagem é que a abordagem generativa consegue gerar textos a partir de uma solicitação ou de um padrão estabelecido.

2.1. Abordagens Discriminativas

As estratégias de Reconhecimento de Entidades Nomeadas (REN) podem ser classificadas conforme dois tipos de abordagem: *token-based* e *span-based*. Na abordagem *token-based*, cada palavra (*token)* do texto é classificada de acordo com os tipos de entidade considerados, além de ser determinado se tal palavra ocorre no início, meio ou fim da denominação da entidade identificada [Finkel et al. 2005, Patil et al. 2020]. Por outro lado, as estratégias *span-based* identificam primeiro todos os *spans* (sequências de *tokens*) menores que um limite dado e, em seguida, classificam cada um deles [Eberts & Ulges 2020, Fu et al. 2021, Liu et al. 2021]. Exemplos de métodos *token-based* que são tradicionalmente utilizados em tarefas REN são aqueles baseados em *Conditional Random Fields* (CRF's) [Finkel et al. 2005, Patil et al. 2020]. CRF's são modelos probabilísticos que utilizam características de um determinado *token* t de um texto (como padrões de letras maiúsculas e minúsculas) e dos tokens adjacentes a t para inferir a categoria de cada *token*.

Entre as estratégias *span-based*, destaca-se o *Span-based Entity and Relation Transformer* (SpERT) [Eberts & Ulges 2020], uma arquitetura neural considerada estado-da-arte na tarefa REN. O SpERT codifica *spans* do texto em uma representação vetorial baseada em modelos pré-treinados tais como o *Bidirectional Encoders from Transformers* (BERT) [Devlin et al. 2019], classificando-os em categorias pré-definidas de entidades ou como "não-entidade". O algoritmo também representa pares de *spans* como entidades no espaço vetorial e os atribui a categorias pré-definidas de relação. Eberts & Ulges [2021] expandiram a arquitetura do SpERT para incluir o agrupamento de menções que

se referem à mesma entidade em diferentes segmentos de um texto. Por fim, Belém et al. [2022] propõem técnicas de reforço contextual e delimitação de entidades baseadas em pré e pós-processamento de dados em documentos oficiais como processos judiciais.

Ji et al. [2020] apresentam uma abordagem para realizar a extração conjunta de entidades e relações em textos. A metodologia proposta utiliza modelos baseados em *transformers*, especificamente o BERT, para codificar os *spans* de texto em representações vetoriais que são usadas para classificar as entidades e suas relações. O artigo destaca a necessidade de uma abordagem *span-based* que leva em consideração o contexto completo em que as entidades aparecem, de modo a superar as limitações das abordagens baseadas em *tokens*. Além disso, é proposto um modelo de atenção baseado em *spans* específicos e representações semânticas contextuais, que permite uma melhor compreensão do contexto das entidades e suas relações.

2.2. Abordagens Generativas

O desempenho de modelos de aprendizado profundo em tarefas como a REN depende diretamente da disponibilidade de dados rotulados para a tarefa em questão. No entanto, a quantidade de dados rotulados pode ser insuficiente para obter um resultado satisfatório. Para contornar essa limitação, abordagens generativas têm sido propostas para gerar novas instâncias de dados que podem ser utilizados para aprimorar o aprendizado discriminativo. Entre os modelos generativos, o BLOOM (*BigScience Large Open-science Open-access Multilingual*) e o ChatGPT (Chat Generative Pre-trained Transformer) surgem como possíveis alternativas para gerar dados sintéticos para a tarefa REN.

Com base na crescente popularidade dos modelos dgenerativos e sua eficácia em diversas tarefas, há um interesse crescente em utilizar esses modelos para gerar dados sintéticos de forma rápida, barata e possivelmente de boa qualidade. No entanto, modelos generativos, como ChatGPT e BLOOM, não são rotuladores de sequência como os modelos para a tarefa REN e podem gerar alucinações, ou seja, rotular entidades que não estão presentes no texto. Para contornar essas limitações, trabalhos como o de Wang et al. [2023] propõem técnicas para aproveitar o poder e tamanho dos *Large Language Models* (LLMs) para utilizá-los como alternativas para os modelos discriminativos. Por outro lado, Tang et al. [2023] analisam a utilização de modelos generativos como ferramentas auxiliares e não como fontes confiáveis para a rotulação de textos clínicos para tarefas REN, questionando a confiabilidade desses modelos em relação à qualidade das respostas geradas e à proteção de dados sensíveis, tendo em vista os textos fornecidos.

Esses estudos indicam uma tendência em explorar o potencial dos modelos generativos como alternativa aos modelos discriminativos, assim como uma forma de realizar o processo de *data augmentation*, permitindo com que tarefas usando modelos REN de domínio específico e com escassez de dados rotulados possam ser aprimorados com um custo relativamente baixo. Apesar de promissoras, as abordagens generativas não constituem ainda substitutos para as discriminativas, devido ao seu elevado custo de infraestrutura e às preocupações com privacidade discutidas anteriormente.

Outra abordagem existente é o *aprendizado ativo*, geralmente utilizado para treinar modelos com poucos dados. No entanto, o tempo e esforço humano necessários para obter a mesma quantidade de dados rotulados seriam significativamente maiores do que os métodos automáticos que empregamos. Portanto, essa abordagem não foi utilizada

neste trabalho.

3. Descrição da Tarefa REN

A tarefa de Reconhecimento de Entidades Nomeadas (REN) consiste em extrair e classificar entidades mencionadas em textos, permitindo que sejam distinguidas entre categorias como Pessoa, Local, Organização, CPF, CNPJ e Número de Telefone, entre outras. Tipicamente, a tarefa REN processa um texto não estruturado, i.e., um texto que não possui indicativos explícitos de quais dos *tokens* que o compõem são entidades nomeadas ou não.

Um exemplo da tarefa REN, dentro do domínio de manifestações de consumidores com foco nas categorias *Organização* e *Produto/Serviço*, seria identificar, dado o texto "Comprei um **telefone móvel** da **Companhia A Ltda.**, mas o alcance do serviço é péssimo. Em outros telefones, isso nunca me aconteceu!" o termo "**telefone móvel**" como uma entidade do tipo *produto ou serviço* e o termo "**Companhia A Ltda.**" como uma entidade do tipo *organização*. A tarefa de extração de relações, apesar de ser tratada pelo método SpERT, não será abordada no escopo deste artigo.

A Tabela 1 apresenta exemplos de reclamações reais realizadas na plataforma **Consumidor.gov.br**. A coluna Reclamação lista os textos das reclamações e a coluna Entidade indica as entidades reclamadas. Nas duas primeiras linhas da tabela temos reclamações referentes a entidades do tipo *organização*, enquanto que nas duas últimas linhas os exemplos de reclamação são de entidades do tipo *Produto/Serviço*.

Tabela 1. Exemplos de reclamações na plataforma consumidor.gov.br. Em negrito a entidade reclamada.

Reclamação	Entidade
Estou a pouco tempo no **mercado livre**, minha primeira venda e já estou com problema. o mercado livre suspendeu a minha conta, sem motivos! preciso que resolvam meu problema o mais rápido possível. indignada com a plataforma **mercado livre**.	Mercado Livre
Estou recebendo cobranças da **Claro** de produtos que não reconheço. Vide tela em anexo. CPF: xxx.xxx.xxx-xx	Claro
Adquiri um **forno elétrico de embutir** diretamente no site da empresa, no dia XXXX, pedido nº XXX , mas o produto apresentou defeito dentro do primeiro ano de uso, reclamei junto a marca mas não tenho sido atendida, há cerca de dois meses estou tentando conseguir uma visita da assistência técnica para identificar e reparar o defeito no produto, o pessoal me liga, mas nunca vem. estou muito decepcionada, meu prejuízo e aborrecimento é incalculável.	Forno elétrico de embutir
Tinha um **plano tim** e me ofertaram outro plano, aí me informaram que iriam me isentar da fatura de junho, que era pra mim ligar pra cancelar meu plano e informar do desconto que me dariam na fatura de junho, quando liguei pra cancelar, me informaram que não seria feito essa solicitação.	Plano TIM

4. Arquitetura Proposta

A Figura 2 apresenta uma visão geral da abordagem PromptNER proposta neste artigo para o reconhecimento de entidades nomeadas. As subseções seguintes detalham cada etapa desta arquitetura.

4.1. Coleta

A coleta de manifestações foi realizada a partir da plataforma **Consumidor.gov.br**, um serviço público que permite a conversa direta entre consumidores e empresas para resolução de conflitos de consumo por meio da Internet. Nessa conversa direta é dispensada a intervenção do Poder Público na tratativa individual, sendo um processo mais rápido e

prático. Vale salientar que todos os dados de reclamações alimentam um banco de dados público, incluindo informações sobre melhores índices de solução e satisfação por empresa.

Por meio da plataforma **Consumidor.gov.br**, foi possível obtermos o conjunto de reclamações sem qualquer marcação das entidades reclamadas no texto. Os dados não estruturados podem ser encontrados na seção Relato do Consumidor da plataforma[5]. As reclamações do consumidor constituem um dado textual que pode auxiliar na caracterização da intenção do consumidor com base em padrões linguísticos recorrentes. Foram coletados 378.574 relatos de consumidores vinculados ao Estado de Minas Gerais[6].

4.2. Prompt-Learning

A segunda parte da arquitetura é similar ao padrão descrito na Seção 1, no qual um conjunto pequeno de exemplos rotulados manualmente é incluído no *prompt*, seguido de um novo exemplo para a qual se busca obter o rótulo produzido pelo modelo generativo, como ilustrado na Figura 3. Para os dados aqui apresentados, foram incluídos quatro exemplos rotulados, seguidos do exemplo não-rotulado para o qual se deseja extrair as entidades.

```
[Texto]: . . .
[produto ou serviço]: Produto A
###
[Texto]: . . .
[produto ou serviço]: Serviço A
. . .
[Texto]: . . .
[produto ou serviço]:
```

Figura 3. Estrutura do prompt utilizado para a tarefa de REN

Nesta etapa, utilizamos o modelo generativo BLOOM, por meio da API de inferência do *huggingface*[7]. A maioria dos LLM's são desenvolvidos por organizações ricas em recursos e são frequentemente mantidos fora do alcance do público. Utilizamos o BLOOM, um modelo de linguagem de acesso aberto com 176B de parâmetros projetados e construído graças à colaboração de centenas de pesquisadores. Especificamente, enviamos uma requisição ao BLOOM de forma similar descrita na Figura 1, obtendo o dado completado pelo rótulo, o qual é então salvo para que possa ser utilizado na etapa seguinte. Aqui, vale observar que acontece um processo de filtragem: rótulos gerados pelo BLOOM que não estejam contidos no texto original da amostra apresentada são descartados e a amostra não é aproveitada. Essa é uma desvantagem do modelo generativo, pois não nos permitiu limitar a geração de *tokens* presentes no texto de entrada.

Uma amostra dos dados coletados é apresentada para os avaliadores para validação da qualidade dos rótulos produzidos pelo BLOOM e filtrados pela automação. Uma vez validados amostralmente, os dados são fornecidos ao modelo SpERT como um conjunto de treino e, após o treinamento (um processo de *fine-tuning* do modelo), as métricas de avaliação são calculadas para aferir a eficácia do fluxo proposto.

[5]`https://consumidor.gov.br/pages/indicador/relatos/abrir`

[6]Tivemos também acesso a milhares de reclamações do Procon-MG, mas como são dados sensíveis, não foi possível utilizá-las neste artigo. Mas elas são parte da motivação para as técnicas propostas.

[7]`https://huggingface.co/inference-api`

4.3. *Fine-Tuning*

O SpERT é um modelo popular e amplamente utilizado na tarefa de identificação de entidades em textos não estruturados. De forma similar a outros *transformers*, ele permite que seja feito o processo de ajuste dos seus pesos (Fine-Tunning), permitindo que o modelo seja adaptado ao domínio do problema, no nosso caso, o reconhecimento de organizações e produtos/serviços a partir de relatos de consumidores. O modelo treinado é salvo, podendo ser utilizado para o reconhecimento de entidades em novos conjuntos de dados sem a necessidade do envio de dados (eventualmente confidenciais) para ambientes externos.

Por fim, mas não menos importante, o processo de treino (fine-tunning) do SpERT pode ser feito de forma local sem transferência de dados, o que é importante para a questão da privacidade e segurança de dados sensíveis, como são por exemplo, aqueles recebidos por meio de órgãos públicos como o Procon.

5. Metodologia de Avaliação

Esta seção descreve a metodologia de avaliação adotada neste artigo. São apresentadas as coleções de dados utilizados (Subseção 5.1), como foi realizada a avaliação por humanos (Seubseção 5.2), as métricas usadas na avaliação experimental (Subseção 5.3) e a parametrização do método SpERT (Subseção 5.4).

5.1. Coleções de Dados

A partir da coleta de 378.574 reclamações (Seção 4.1), 7.858 dessas reclamações foram rotuladas pelo modelo generativo BLOOM. A Tabela 2 apresenta a quantidade de reclamações para entidades dos tipos *Organização* e *Produto ou Serviço* rotuladas pelo BLOOM. Após o processo de rotulação automática, estes dados como entrada para SpERT.

Tabela 2. Estatísticas da coleção de dados consumidor.gov.br

Tipo de Entidade	Quantidade de reclamações
Organização	3129
Produto ou Serviço	4729

Para avaliação, foi utilizado o processo de 5-*fold cross validation*, no qual em cada *fold* os dados são divididos em 5 partições. Três dessas partições constituem o *conjunto de treino* que contém, tanto os poucos exemplos que precisam ser rotulados manualmente, como os exemplos rotulados automaticamente por meio do prompt no LLM. Esses dados rotulados são usados para ajustar os pesos do modelo (*fine-tuning*). Uma outra partição é usada como *conjunto de validação* para parametrização do modelo. A última partição, o *conjunto de teste*, consiste dos dados aos quais o REN é aplicado para ser avaliado.

Nossa metodologia compara cenários onde é possível que apenas poucos dados sejam rotulados manualmente (Cenário-1) em contraste com o cenário onde temos dados rotulados a partir de *promp leaning* (Cenário-2). Utilizamos 100 manifestações (a totalidade de dados manualmente rotulados disponível) como treinamento no Cenário-1 e 7.858 no Cenário-2. Nos dois cenários, mantivemos o mesmo conjunto de teste.

5.2. Avaliação por Humanos

A partir das reclamações rotuladas através de nossa abordagem baseada em *prompts* com base no LLM BLOOM, fizemos uma inspeção manual com três avaliadores analisando

uma amostra de 100 reclamações, sendo 50 relacionadas a entidades do tipo *Organização* e 50 do tipo *Produto/Serviço*. Fornecemos para cada avaliador uma planilha contendo as 100 reclamações para que fosse avaliado se o modelo generativo conseguiu identificar a entidade reclamada, de modo que um avaliador não tivesse acesso à resposta dos demais. Vale ressaltar que esta é uma tarefa custosa para uma pequena equipe de trabalho, principalmente diante da escrita livre que os consumidores utilizam.

5.3. Métricas de Avaliação

Para avaliação dos resultados foram utilizadas as métricas Precisão, Revocação e F1, que capturam diferentes aspectos da eficácia do reconhecimento de entidades. Considerando x um tipo de entidade (e.g., x = Pessoa ou x = Organização), a precisão do algoritmo para reconhecer entidades do tipo x é calculada pelo número de acertos em relação ao total de vezes que o algoritmo reconheceu o tipo x. Já a Revocação mostra o quanto o algoritmo conseguiu cobrir as menções a entidades de um tipo x. Por fim, a métrica F1(x) é definida como a média harmônica entre a Precisão(x) e a Revocação(x). Em nossos resultados, apresentamos a medida F1 por categoria de entidade e a média entra elas (Macro-F1).

5.4. Parametrização do SpERT

Para parametrização da estratégia SpERT, foram utilizados os valores recomendados pelos seus autores [Eberts & Ulges 2020]: taxa de aprendizado definida como $l_r = 5 \times 10^{-5}$, número de épocas $t = 20$, número de exemplos negativos por frase n^- = 100 (tanto para entidades como para relações) e tamanho de cada *batch* b_s = 2.

6. Resultados Experimentais

Esta seção apresenta os resultados experimentais que visam responder às três questões de pesquisa colocadas.

6.1. QP1: Qual o nível de concordância entre os avaliadores sobre a rotulação de reclamações usando LLMs?

Na amostra das 50 reclamações da entidade Organização, os três avaliadores concordaram que o *BLOOM* conseguiu identificar a respectiva organização para 43 delas. No caso do reconhecimento da entidade Produto/Serviço, das 50 reclamações, os três avaliadores concordaram que o *BLOOM* conseguiu identificar no texto o Produto/Serviço correspondente a 28 delas.

Por fim, utilizamos o coeficiente alfa de Krippendorff, que mede o nível de concordância entre os avaliadores, sendo amplamente utilizado quando a quantidade de avaliadores é maior que dois [Akter & Wamba 2016, Fabbri et al. 2021]. A escala desta métrica varia de -1 (máxima discordância) a 1 (unanimidade).

O coeficiente alfa de Krippendorff obtido em nossa análise foi de 0.53, o que sugere uma concordância entre os avaliadores suficiente para validar os resultados, mas ainda distante de uma unanimidade (representada pelo valor 1), sugerindo que a tarefa de rotulação manual é complexa e sujeita a erros [Zhu et al. 2023]. Isso motiva soluções que dependam de menos dados rotulados manualmente, como é o caso da abordagem PromptNER proposta.

6.2. QP2: Qual é a eficácia da rotulação obtida a partir de LLMs como o BLOOM?

A Tabela 3 apresenta a eficácia do BLOOM nas métricas de precisão, revocação e F-score por tipo de entidade de uma amostra contendo 100 reclamações e considerando as avaliações de três avaliadores. O BLOOM alcança uma F-score de 0,83 para a entidade Organização e 0,56 para produto e serviço, resultando em uma Macro-F1 de 0,695.

Podemos observar uma eficácia superior em relação à entidade organização em comparação a Produto/Serviço, porque os termos associados a organização ocorrem em um contexto mais definido como seguido por preposições tais como "ao", "de" e "em". Outro argumento é que a identificação de produto/serviço em alguns casos demanda conhecimento, também, da organização associada, de forma a conhecer os serviços por ela disponibilizados. No prompt de exemplo trazido na Figura 4, é preciso conhecer a organização **Correios** para identificar o termo "logística reversa" como um serviço oferecido. Um último ponto é que o conjunto de entidades do tipo organização é mais restrito do que o de produto/serviço, fazendo com que o modelo BLOOM tenha mais exemplos no seu processo de treinamento.

```
[Texto]: Fui ao Correios e não consegui utilizar a logística reversa para devolver meu
Iphone-X de volta à Apple
[Produto ou Serviço]: logística reversa
[Produto ou Serviço]: Iphone-X
[Organização]: Correios
[Organização]: Apple
```

Figura 4. Exemplo de um *prompt* com organização, produto e serviço.

Tabela 3. Resultado da Precisão, Revocação e F-Score na amostra do BLOOM

Tipo	Precisão	Revocação	F-Score
Organização	0.86	0.80	0.83
Produto ou Serviço	0.56	0.56	0.56

6.3. QP3: Qual é a eficácia dos modelos do estado-da-arte como o SpERT em dados rotulados a partir de LLMs?

Como *baseline*, utilizamos o modelo SpERT utilizando 100 exemplos inspecionados manualmente, representando a típica situação de escassez de dados rotulados manualmente, sem a utilização da rotulação automática baseada em LLMs. Nossa arquitetura PromptNER, por outro lado, utiliza a rotulação do modelo generativo, que obtém uma quantidade maior de dados rotulados a um custo muito menor que o da rotulação manual. Ambos os modelos estão utilizando a mesma partição de teste do *5-fold cross validation*, para garatir que são avaliados o mesmo conjunto de reclamação, a diferença está na partição do treino.

Apresentamos os resultados do baseline e do PromptNER na Tabela 4. Na tarefa de reconhecer a entidade organização, a arquitetura PromptNER conseguiu uma melhoria de Precisão de 41,0% (0.39 vs. 0.55), 37% na Revocação (0.51 vs. 0.70) e 41% da F1 (0.44 vs. 0.62) em comparação com SpERT, o que mostra a importância da utilização do modelo generativo na rotulação de dados e por consequência melhoria do modelo classificativo treinado. No reconhecimento de produtos/serviços os resultados foram mais impactantes, conseguindo alcançar uma melhoria em precisão, revocação e f-score de 268,7% (0.16 vs. 0.43), 80,0% (0.30 vs. 0.54) e 128,6% (0.21 vs. 0.48), respectivamente. Outro ponto que podemos observar na Tabela 4 é que o intervalo de confiança é pequeno

em relação a média, o que mostra que o resultado foi estável entre as partições (folds), e portanto o modelo é bem generalizável (i.e., robusto) para diferentes partições de treino e teste.

A Tabela 5 apresenta exemplos de acertos e erros do SpERT e PromptNER em relação ao reconhecimento de Produtos/Serviços. Nas linha 1 e 2, apenas o PromptNER acerta o produto reclamado (em azul). Particularmente, na linha 1 o SpERT sugeriu parcialmente o nome da organização "bahia" (relativo a Casas Bahia), que não é o foco da tarefa de reconhecimento de produtos. Já na linha 2, o SpERT sugeriu o termo "celular", que não é produto reclamado. Na linha 3, temos uma reclamação em que ambos os modelo erram – ambos sugerem a organização em vermelho – o produto reclamado correto está em azul. Percebemos a complexidade dessa reclamação – seria preciso que o modelo relacionasse o termo "contestação" com "faturas em aberto" para que o reconhecimento seja feito de forma adequada.

Tabela 4. Resultado Precisão, Revocação e F-Score com intervalo de 95% de confiança do modelo SpERT e da arquitetura PromptNER

Método	Entidade	Precisão	Revocação	F-Score
SpERT	Organização	0.39 ± 0.09	0.51 ± 0.04	0.44 ± 0.05
PromptNER	Organização	0.55 ± 0.03	0.70 ± 0.02	0.62 ± 0.02
SpERT	Produto ou Serviço	0.16 ± 0.04	0.30 ± 0.04	0.21 ± 0.03
PromptNER	Produto ou Serviço	0.43 ± 0.02	0.54 ± 0.03	0.48 ± 0.02

Tabela 5. Exemplos de acertos e erros do PromptNER e SpERT.

Reclamação
Bom dia, eu fiz uma compra nas casas bahia de 2 produtos um kit de liquidificador, batedeira e espremedor de laranja e uma escova rotativa e cancelei o pedido pela demora da entrega, pois eles me deram uma data e não foi cumprida ... pedi o estorno do valor, eles me estornaram o valor do kit e não da escova rotativa, e no site consta que recebi a escova. entrei em contato com eles diversas vezes, ate pedi a comprovação que eu tinha recebido, assinatura de quem recebeu o produto. Eles falam que vão estornar o valor e ate hoje nada. eu já paguei todas as parcelas
em 2018 contratei um plano pós-pago para o celular número xxx , no valor de 49,90. De uns meses para cá, o valor do plano vem recebendo aumento, chegando atualmente ao valor de r$ 67,00. Nas faturas atuais, vêm cobrando sva 's (serviço adicionais no plano) que não contratei, não utilizo e não me foram informados na hora da contratação, e que aumentam drasticamente o valor da fatura. liguei dia 10/03/2021 na tim e falei com uma atendente, que não soube me explicar o porque dos valores das faturas estarem aumentando e não resolveu nada sobre minha situação
fui até uma loja da operadora tim, contratar um plano, mas a atendente me informou que havia duas faturas em aberto em meu cpf, informei a mesma que não conheço o numero preenchi uma carta a próprio punho e encaminhei pra operadora pedindo a contestação dos valores em aberto e informando desconhecer a linha, mas até hoje nada foi resolvido.

7. Conclusões e Trabalhos Futuros

Neste artigo, tratamos o problema de REN em dados reais de reclamações coletadas do sitio **Consumidor.gov.br**, com o objetivo de identificar Organizações e Produtos/Serviços mencionados no texto. Para isso usamos a abordagem PromptNER que consiste em duas etapas: (1) rotulação automática de dados públicos baseada em LLMs, e (2) aplicação dos dados rotulados como treino para modelos escaláveis e que podem ser mantidos em uma infraestrutura mais simples e privativa, sem a necessidade de intercâmbio de dados, eliminando custos e a preocupação com a submissão de dados confidenciais para os ambientes externos dos LLMs. Nosso modelo alcançou resultados promissores, com ganhos de 41% até 129% em F-Score em relação ao modelo do estado-da-arte (SpERT) treinado com dados rotulados manualmente, com a vantagem adicional de dispensar o custoso processo de rotulação manual. Em trabalhos futuros, utilizaremos o ChatGPT como LLM, na avaliação de diferentes *prompts* para REN e auto-aprendizado (*self-learning*).

Agradecimentos

Gostaríamos de agradecer ao Ministério Público de Minas Gerais, por meio do projeto Capacidades Analíticas, pelo apoio institucional a este trabalho, bem como à AWS, CAPES, CNPq, FAPESP e FAPEMIG pelo apoio individual aos seus autores.

Referências

Akter, S. & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2):173–194.

Belém, F., Ganem, M., França, C., Carvalho, M., Laender, A., & Gonçalves, M. (2022). Reforço e Delimitação Contextual para Reconhecimento de Entidades e Relações em Documentos Oficiais. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 292–303.

Brunner, U. & Stockinger, K. (2020). Entity Matching with Transformer Architectures - A Step Forward in Data Integration. In *Proceedings of the International Conference on Extending Database Technology*, pages 463–473.

Caputo, A., Basile, P., & Semeraro, G. (2009). Boosting a Semantic Search Engine by Named Entities. In *Foundations of Intelligent Systems*, pages 241–250.

de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., & Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *Information Processing & Management*, 60(4):103336.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Eberts, M. & Ulges, A. (2020). Span-based Joint Entity and Relation Extraction with Transformer Pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 2006–2013.

Eberts, M. & Ulges, A. (2021). An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3650–3660.

Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., & Radev, D. R. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370.

Fu, J., Huang, X., & Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Annual Meeting of the Association for Computational Linguistics*, pages 7183–7195.

Ji, B., Yu, J., Li, S., Ma, J., Wu, Q., Tan, Y., & Liu, H. (2020). Span-based Joint Entity and Relation Extraction with Attention-based Span-specific and Contextual Semantic Re-

presentations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99.

Liu, C., Fan, H., & Liu, J. (2021). Span-Based Nested Named Entity Recognition with Pretrained Language Model. In Jensen, C. S., Lim, E.-P., Yang, D.-N., Lee, W.-C., Tseng, V. S., Kalogeraki, V., Huang, J.-W., & Shen, C.-Y., editors, *In Processing of the 26th International Conference Database Systems for Advanced Applications*, pages 620–628.

Luo, X., Xue, Y., Xing, Z., & Sun, J. (2022). PRCBERT: Prompt Learning for Requirement Classification using BERT-based Pretrained Language Models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Mangaravite, V., Carvalho, M., Cantelli, L., Ponce, L. M., Campoi, B., Nunes, G., Laender, A. H. F., & Gonçalves, M. A. (2022). DedupeGov: Uma Plataforma para Integração de Grandes Volumes de Dados de Pessoas Físicas e Jurídicas em Âmbito Governamental. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 90–102.

Niu, F., Zhang, C., Ré, C., & Shavlik, J. W. (2012). DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. In *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*, pages 25–28.

Patil, N., Patil, A., & Pawar, B. (2020). Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188. International Conference on Computational Intelligence and Data Science.

Silva, L., Canalle, G. K., Salgado, A. C., Lóscio, B., & Moro, M. (2019). Uma Análise Experimental do Impacto da Seleção de Atributos em Processos de Resolução de Entidades. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 37–48.

Silva, R. M., Gomes, G. C. M., Alvim, M. S., & Gonçalves, M. A. (2022). How to build high quality L2R training data: Unsupervised compression-based selective sampling for learning to rank. *Information Sciences*, 601:90–113.

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). Does synthetic data generation of llms help clinical text mining? *Computer Science Archive*, abs/2303.04360.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. *Computer Science Archive*, abs/2304.10428.

Ye, F., Huang, L., Liang, S., & Chi, K. (2023). Decomposed Two-Stage Prompt Learning for Few-Shot Named Entity Recognition. *Information*, 14(5).

Zhu, Y., Ye, Y., Li, M., Zhang, J., & Wu, O. (2023). Investigating annotation noise for named entity recognition. *Neural Comput. Appl.*, 35(1):993–1007.