A oclusão pode facilitar a compreensão humana? Avaliação de explicabilidade no reconhecimento de entidades nomeadas

Alexandre Augusto Aguiar Gomes[1], Leonidas J. F. Braga[1], Marcos P. C. Azevedo[1], Gabriel Assunção[2], Arthur Carvalho[3], Michele A. Brandão[4], Daniel H. Dalip[1], Flávio Cardeal Pádua [1],

[1]Centro Federal de Educação Tecnológica de Minas Gerais
Belo Horizonte, MG – Brasil

[2]Universidade Federal de Minas Gerais
Belo Horizonte, MG – Brasil

[3]Blip
Belo Horizonte, MG – Brasil

[4]Instituto Federal de Minas Gerais (IFMG)
Ribeirão das Neves, MG – Brasil

{aaugustoag, lebragacf, marcosazevedo2112}@gmail.com,
arthurc@blip.ai, gabrieloa@ufmg.br,
michele.brandao@ifmg.edu.br, {hasan, cardeal}@cefetmg.br

Abstract. *Explainability techniques help users understand the results of a machine learning model. This work investigates whether the Occlusion explainability technique can generate responses similar to those expected by humans in word classification for Named Entity Recognition. For this, we use a bidirectional LSTM, the CoNLL 2003 dataset, and the manual annotation of 849 sentences, thus creating a reference database. The results show that Occlusion is capable of indicating at least one word that is relevant and compatible with human understanding.*

Resumo. *Técnicas de Explicabilidade são métodos que auxiliam usuários a entender os resultados de um modelo de aprendizado de máquina. Nesse contexto, este trabalho investiga se a técnica de explicabilidade de Oclusão consegue gerar respostas similares às esperadas por humanos na classificação de palavras para o Reconhecimento de Entidades Nomeadas. Para isso, utilizou-se uma LSTM bidirecional e o conjunto de dados CoNLL 2003, bem como foi utilizado a anotação manual de 849 sentenças criando-se, assim, uma base de dados de referência. Os resultados mostram que a Oclusão é capaz de indicar pelo menos uma palavra relevante e compatível com a compreensão humana.*

1. Introdução

A Explicabilidade em Inteligência Artificial (XAI, do inglês, *Explainable Artificial Intelegence*) é uma ferramenta utilizada para entender os motivos pelos quais o modelo de Aprendizado de Máquina (ML, do inglês, *Machine Learning*) realizou uma determinada predição [Van Lent et al. 2004]. Elas vêm sendo aplicadas nos mais diversos problemas, desde Visão Computacional ao Processamento de Linguagem Natural (NLP, do inglês,

Natural Language Processing). Em NLP, tais técnicas podem permitir ao usuário identificar quais palavras são importantes para a compreensão de uma sentença ou visualizar quais foram as palavras relevantes para efetuar uma determinada predição [Hu 2018]. Por exemplo, em tarefas de identificação automática de classes gramaticais ou Reconhecimento de Entidades Nomeadas (NER, do inglês, *Named Entity Recognition*).

O Reconhecimento de Entidades Nomeadas aplicado em NLP objetiva identificar palavras ou expressões de forma que seja mais fácil entender e processá-los. A ambiguidade de palavras na utilização de NER é um dos principais desafios nessa área sendo, geralmente, utilizadas as Redes Neurais Artificiais e arquiteturas do tipo biLSTM (do inglês *Bidirectional Long Short - Terms Memory*) por apresentarem uma precisão consideravelmente alta (acima de $\sim 90\%$ [Vajjala and Balasubramaniam 2022]).

A explicabilidade pode prover ao usuário explicações de uma determinada classificação, permitindo uma melhor tomada de decisão e cumprir, inclusive, legislações de Proteção de Dados [Goodman and Flaxman 2017]. Além disso, por meio delas, pode-se perceber tendências discriminatórias [Pedreshi et al. 2008] e mitigar problemas causados por falhas de predição [Guidotti et al. 2018]. Em NER, pode-se analisar o impacto que as demais palavras da frase tem na nomeação da entidade a ser classificada.

Para classificar as entidades, este trabalho utiliza uma LSTM bidirecional e o conjunto de dados CoNLL 2003, amplamente utilizados na literatura. Aqui, também é investigada a precisão da técnica de Explicabilidade (Oclusão) por meio da sua capacidade em gerar respostas similares às esperadas por humanos. Para realizar essa análise, foi utilizada a anotação manual, realizada por meio de uma aplicação Web (uma contribuição deste trabalho), de um conjunto de 849 sentenças, criando-se, assim, uma base de dados de referência. Até onde se sabe, não há na literatura análises do uso da Oclusão em NER.

Este trabalho está organizado da seguinte forma. A Seção 2 discute sobre os trabalhos relacionados na área de NLP e XAI; a Seção 3 descreve as principais etapas da metodologia; a Seção 4 descreve os resultados quanto à precisão da Oclusão, uma análise dos valores de relevância de uma palavra para humanos e os impactos da qualidade do modelo de ML nos resultados; e a Seção 5 apresenta as considerações finais, desafios e trabalhos futuros. Seções estas, nas quais, buscou-se responder às perguntas de pesquisa: A Oclusão é uma alternativa para se obter uma explicação para uma predição em NER? Qual o nível de confiabilidade para essa tarefa? Quais as limitações para essa técnica? E quais as possibilidades de estudos futuros?

2. Trabalhos Relacionados

Em NLP, com as ferramentas de busca em larga escala na *web*, estamos sendo confrontados com um volume cada vez maior de informações em texto [Brin and Page 1998], mas que, nem sempre, podemos compreender seu valor potencialmente escondido [Siddharthan 2002]. Esse fenômeno impulsionou pesquisas nas diversas áreas no sentido de melhorar a predição dos modelos, como Análise de Sentimentos [Yang et al. 2019] e NER [Vajjala and Balasubramaniam 2022], tal que NER vem se destacando devido ao seu uso em Recuperação de Informação *web* [Liu et al. 2022].

Em seu trabalho, [Vajjala and Balasubramaniam 2022] apresentam as principais arquiteturas do estado da arte em NER. Um aspecto comum a essas arquiteturas são suas complexas redes cujas repostas são praticamente impostas ao usuário final

[Adadi and Berrada 2018], motivando agências governamentais a aprovar leis de Proteção de Dados (GDPR, do inglês, *General Data Protection Regulation*) que busquem dar transparência nas tomadas de decisões automatizadas, provendo ao usuário direito de obter explicações significativas sobre as lógicas de decisão [Goodman and Flaxman 2017].

Um risco inerente ao uso de tecnologias de predição automática é a possibilidade de tomar decisões equivocadas, podendo levar, inclusive, a fatalidades, como acontecido em março de 2018 no Arizona (EUA)[1] [Guidotti et al. 2018]. Além disso, [Pedreshi et al. 2008] observou, com o uso de Mineração de Dados, que decisões baseadas em dados históricos, podem refletir tendências discriminatórias, que podem ser passadas aos modelos de ML durante o treinamento. Tais problemas poderiam ser mitigados ou resolvidos com a utilização de XAIs seguindo a metodologia de justificar, controlar, melhorar e descobrir, conforme a explicação dada à predição, evitando e/ou corrigindo eventuais falhas do modelo ou vícios presentes no banco de dados [Adadi and Berrada 2018].

Algumas técnicas para gerar explicações em problemas de NLP envolvem engenharia reversa como LRP (do inglês, *Layer-Wise Relevance Propagation*) [Arras et al. 2017, Hu 2018] e Perturbação da Vizinhança utilizando análise de saliência por derivada [Wallace et al. 2018]. Em particular, [Arras et al. 2017] implementa uma LRP com aplicação em Análise de Sentimentos de textos. Seu trabalho motivou [Hu 2018] a testar essa abordagem em NER. Apesar de ambas as tarefas pertencerem ao conjunto de NLP, suas arquiteturas diferem no que consiste em dimensões das saídas uma vez que as tarefas de Análise de Sentimentos são n entradas para 1 saída e NERs são n entradas para n saídas.

A utilização da Oclusão [Robnik-Šikonja and Kononenko 2008] em NLP surgiu como suporte para Modelagem de Linguagem com o uso de Axiomas em tarefas supervisionadas, permitindo a substituição de palavras que não estão no vocabulário [Harbecke 2021]. A adaptação de [Hu 2018] para o uso de LRP em NER e a abordagem de [Harbecke 2021] no uso da Oclusão serviram de motivação para a aplicação desta técnica em NER neste trabalho. Soma-se à isso, o fato da Oclusão ser plenamente explicável à um Humano, respeitando-se assim regras de Proteção de Dados, mesmo na XAI. Ainda existem poucas referências na literatura abordando as possibilidades da aplicação de XAI em NER ou NLP [Danilevsky et al. 2020]. Dada a crescente demanda por técnicas que permitem explicar os modelos cada vez mais complexos de IA e ML, este trabalho busca contribuir com uma análise do uso da Oclusão aplicada em NER.

3. Metodologia

Esta seção apresenta a metodologia para analisar a Explicabilidade por Oclusão aplicada em NER, conforme sumariza a Figura 1. Inicialmente, treinou-se um modelo biLSTM para identificação das entidades nomeadas. Tal modelo foi criado considerando uma amostra do conjunto de dados (treino). Em seguida, foi selecionada uma amostra de teste para realizar a explicabilidade e sua análise. Aplicou-se a essa amostra o modelo biLSTM com objetivo de obter as predições das entidades nomeadas. Por meio das predições foram realizadas explicações utilizando a Oclusão. Com o objetivo de analisar os resultados da Oclusão, foi feito uma rotulação manual indicando, para uma palavra alvo na frase,

[1] `https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html`.

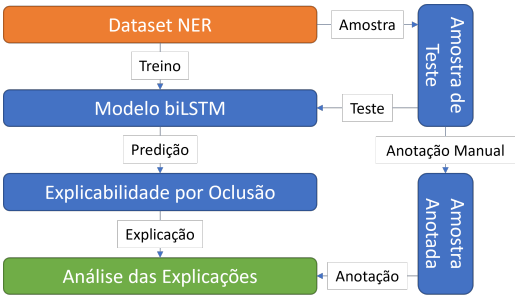


Figura 1. Fluxo da Metodologia

Tabela 1. Distribuição do CoNLL 2003 em inglês [Sang and De Meulder 2003].

Dataset	Frases	Palavras	PES	LOC	ORG	MIS
Training set	14.987	203.621	6.600	7.140	6.321	3.438
Development set	3.466	51.362	1.842	1.837	1.341	922
Test set	3.684	46.435	1.617	1.668	1.661	702

quais eram as palavras relevantes para classificá-la em um determinado rótulo. Assim, foram realizadas análises considerando as anotações e as relevâncias obtidas por meio da explicação. O restante desta seção descreve em detalhes cada uma dessas etapas.

Base de Dados de NER. O dataset CoNLL 2003 [Sang and De Meulder 2003] é um conjunto de dados em inglês e alemão, com entidades nomeadas que possuem os seguintes rótulos **Pessoas**, **Localidades**, **Organizações** e **Miscelânea**, e consiste de frases retiradas do Reuters Corpus [Russell-Rose et al. 2002]. Esse corpus é formado por notícias da Reuters entre agosto de 1996 e agosto de 1997 possuindo 301.418 tokens em 1.393 artigos. Esse dataset foi escolhido para o desenvolvimento do presente trabalho devido à sua ampla utilização em pesquisas de NER. O site *paperwithCode*[2] apresenta pesquisas que utilizam esse conjunto de dados tais como [Wang et al. 2020], que usou arquiteturas LSTM e Transformers, [Schweter and Akbik 2020](Transformers) e [Yamada et al. 2020] (Transformers), esses com valores de Precisão próximos de 94%. Devido à quantidade de referências em inglês para Explicabilidade e sendo o CoNLL2003 um dos principais conjuntos de dados para NER, optou-se por usá-lo em detrimento de outros em português.

Modelo de Aprendizado de Máquina (biLSTM). Este trabalho baseou-se na implementação de [Hu 2018] que utilizou a XAI LRP aplicada para NER por meio de um modelo com a arquitetura biLSTM. A representação textual utilizada foi o *GloVe Embeddings* [Pennington et al. 2014] e CoNLL 2003 Dataset [Sang and De Meulder 2003] no treinamento. Cinco modelos diferentes foram treinados, aumentando gradualmente o número de épocas. Em uma rede neural, em cada época os pesos do modelo são ajustados de acordo com uma função de otimização[3], retornando a perda (*loss*) e a acurácia da mesma. Quanto mais épocas de treinamento, mais o modelo será ajustado ao treino. Isso foi realizado para avaliar o impacto do método de explicabilidade em diferentes modelos. Considerando o desempenho de cada modelo em prever entidades nomeadas, a Tabela 2 apresenta os valores da função de *Loss* e acurácia do treino e da validação. Como pode-se observar, os modelos apresentaram uma taxa de acurácia alta (acima dos 95%) tanto

[2]`https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003`

[3]A função de otimização utilizada nos treinos foi a Entropia Cruzada.

Tabela 2. Modelos utilizados para avaliar as explicações.

Modelo	-	Épocas	Treinamento		Validação	
			Loss	Acurácia	Loss	Acurácia
Modelo 1	-	5 Épocas	0.0301	99.13%	0.0348	99.00%
Modelo 2	-	10 Épocas	0.0194	99.42%	0.0252	99.27%
Modelo 3	-	20 Épocas	0.0109	99.66%	0.0230	99.35%
Modelo 4	-	50 Épocas	0.0036	99.88%	0.0297	99.38%
Modelo 5	-	100 Épocas	0.0020	99.93%	0.0365	99.38%

no treino quanto na validação. Portanto, esses modelos foram utilizados para realizar as predições das entidades nomeadas nas análises deste trabalho.

Implementação da Oclusão. Este trabalho tem por objetivo avaliar a explicação do modelo e a fidelidade do explicador ao modelo de ML utilizado. Para tal, foi implementada uma variação da técnica de Oclusão proposta por [Robnik-Šikonja and Kononenko 2008], usada em explicabilidade de imagens, para a aplicação em NER. Essa implementação remove cada palavra da sentença medindo a variação da predição para rotulação da palavra alvo na saída, conforme mostra o Algoritmo 1.

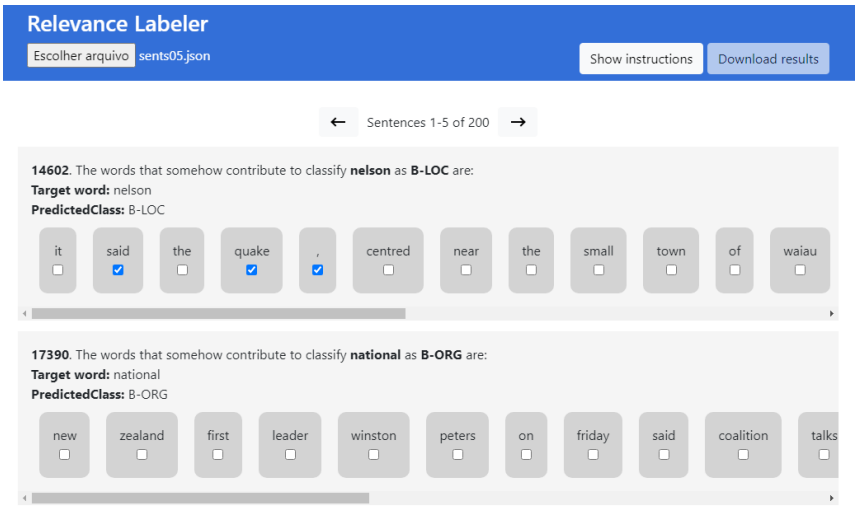
Algoritmo 1 Técnica de Oclusão aplicada em NER.

```
Entrada: modeloTreinado, frase, palavraAlvo
Saída: relevância
início
    inicialização;
    predição(p) = modeloTreinado(frase, palavraAlvo);
    relevância = [];
    repita
        if p != i: then
            frase(i) = frase.remove(palavra(i));
            predição(i) = modeloTreinado(frase(i), palavraAlvo);
            relevância(i) = ||predição(p) − predição(i)||;
        relevância += relevância(i);
    até para cada palavra(i) na frase;
fim
```

Dado uma frase S composta por n palavras $\{w_1, w_2, ..., w_n\}$ e uma determinada palavra $w_k \in S$ que desejamos uma explicação de sua classificação. Assumimos que as demais palavras da frase influenciam na classificação de w_k. Assim, a saída deste método de explicabilidade será um ranking considerando as demais palavras $w_i \in S$ ordenado pela sua relevância r_i ao classificar w_k. Para gerar tal relevância, a Oclusão analisa o impacto na probabilidade de predição da classe prevista em w_k ao remover cada palavra w_i da frase. Especificamente, para cada palavra w_i, criamos outra frase S'_i, sem w_i, e prevemos a classe de w_k em S'_i. Assim, a relevância r_i de cada palavra w_i é calculado da seguinte forma $r_i = |p_k - p'_{ik}|$, em que p_k e p'_{ik} são as probabilidades ao prever a classe da palavra w_k nas frases S e S'_i.

Análise dos Resultados. A Oclusão gera um ranking de palavras relevantes ordenada por valores entre 0 e 1. Este trabalho analisa tais valores de relevância R e também como essas explicações se aproximam da explicação feita por humanos. Para isso, foi feita uma anotação manual dessa relevância e utilizadas métricas de avaliação.

Figura 2. Rotulador *Web* desenvolvido no escopo deste trabalho.



Amostra Anotada Manualmente com Palavras Relevantes. Foram selecionadas aleatoriamente 849 frases da base de dados CoNLL 2003. Os requisitos para se selecionar uma frase são: *(i)* possuir mais que cinco palavras e menos que cinquenta; e *(ii)* possuir, ao menos, uma palavra rotulada com uma das categorias, tal que sua distribuição foi: 126 Palavras rotuladas como PER (Pessoa), 420 como LOC (Localidade) e 303 como ORG (Organização). Esses dados foram anotados com o auxílio de voluntários, que pertencem a grupos de estudantes de nível técnico, graduação e mestrado tanto de áreas de exatas, como tecnologia da informação e humanas. Esses voluntários decidiram quais as palavras relevantes para a rotulação da palavra alvo em cada frase da amostra. Elas foram agrupadas em um conjunto de relevância não possuindo quantidade máxima, podendo, inclusive, ser um conjunto vazio. Para a rotulação, foi desenvolvida uma aplicação Web, conforme mostra a Figura 2, na qual basta o usuário selecionar as caixa consideras relevantes.

Como o conceito de relevância é subjetivo, antes da classificação foram apresentados exemplos do que pode ser considerado relevante no contexto deste trabalho da seguinte forma: *(i)* Palavras que possuem correlações gramaticais como concordâncias verbais e nominais – por exemplo, o verbo 'ir' indica que uma Pessoa vai para algum lugar (Local ou Organização), e o adjetivo 'inteligente' é fortemente relacionado com Pessoa; *(ii)* Palavras com interdependências como nomes e sobrenomes – por exemplo, se o Nome é de um Local, o Sobrenome também é de um Local e se existe um Sobrenome, necessariamente deve existir um Nome. Entende-se por Nome, no contexto desse trabalho, os rótulos com prefixo B- (do inglês *begin*, inicial) e como sobrenome os rótulos com prefixo I- (do inglês *Inner*, 'do meio'); e *(iii)* Outras tal qual o voluntário entender, conforme contexto, que colaboraram no entendimento da rotulação.

Métrica de Avaliação. Para calcular a precisão da ordenação obtida pela técnica de explicabilidade, foi calculada a média da precisão nas k_esimas posições ($\overline{P@k}$) considerando todas as frases rotuladas: $\overline{P@k} = \frac{\sum_{i=0}^{N} P@k_i}{N}$, em que N é a quantidade de frases e $P@k_i$ representa a Precisão da Explicação para i_esima frase para as primeiras k posições. Tal precisão é calculada como a proporção da quantidade de palavras relevantes até a k_esima posição do ranking.

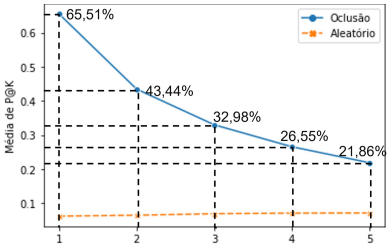


Figura 3. Precisão na k-ésimas posições.

4. Resultados

4.1. Precisão da Explicação por Oclusão considerando a Explicação Humana

Para avaliar se a Oclusão cumpre sua função de prover uma explicação miníma condizente com a explicação por humanos, foi utilizada a Precisão nas k-ésimas posições do rank ($P@k$). A Figura 3 apresenta, pra cada posição k a precisão média das explicações ($\overline{P@k}$), cf. Eq. 1. Como método de referencia, foi criado uma técnica na qual as palavras são ordenadas de forma aleatórias que seria o pior caso possível, de forma a permitir avaliar se a XAI está produzindo resultados melhores que um resultado aleatório.

$$\overline{P@k} = \frac{\sum_{i=0}^{N} P@k_i}{N} \tag{1}$$

Pode-se verificar na Figura 3 que a $\overline{P@k}$ diminui à medida que se aumenta o valor de k. Esse comportamento reflete o aumento da quantidade das palavras testadas e, com isso, o aumento na probabilidade de erro. Analisando os valores de relevância, percebeu-se que a Oclusão detecta geralmente apenas uma palavra com valor de Relevância R acima de 10^{-2}. Em outras palavras, ela é capaz de identificar qual é a palavra mais relevante, porém, pode falhar em identificar as demais. Uma análise mais aprofundada sobre esse comportamento será discutido na Seção 4.2. Na Figura 3, a precisão para k entre 1 e 3 permanece acima de 33%, o que significa que, em média, a Oclusão aponta pelo menos uma palavra relevante em um ranking das três palavras mais relevantes na frase. Já para k igual a 4, acima de 25% e k igual a 5, acima de 20%.

4.2. Análise dos Valores de Relevância Estimados pela Oclusão

Esta seção analisa o valor da Relevância e compara esse valor com a relevância apontada por um humano na anotação manual. Além disso, esta seção visa compreender quando pode-se considerar o valor de Relevância R, estimado pela Oclusão, suficiente para uma palavra ser considerada Relevante. A escolha desse valor é subjetiva e pode variar com cada problema a depender dos próprios parâmetros do problema ou do modelo, como os valores da predição y. A Figura 4 levou em consideração a distribuição da relevância das 5 palavras mais bem ranqueadas pela Oclusão em cada frase. Já na Figura $4(a)$, considerou-se apenas as palavras Relevantes (anotadas pelos voluntários) e na Figura $4(b)$, as não relevantes (não anotadas pelos voluntários). Os valores de relevância desses gráficos estão expressos por ordem de grandeza na escala logarítmica.

Percebe-se, na Figura $4(a)$ que a ordem de grandeza acima de 10^{-1} possui aproximadamente $37,54\%$ palavras relevantes anotadas, e acima de 10^{-2} aproximadamente

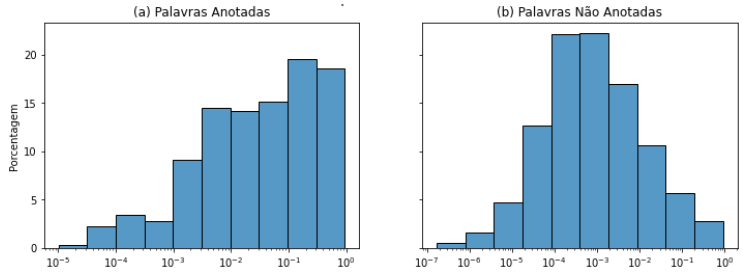


Figura 4. Top 5 palavras relevantes anotadas por ordem de grandeza.

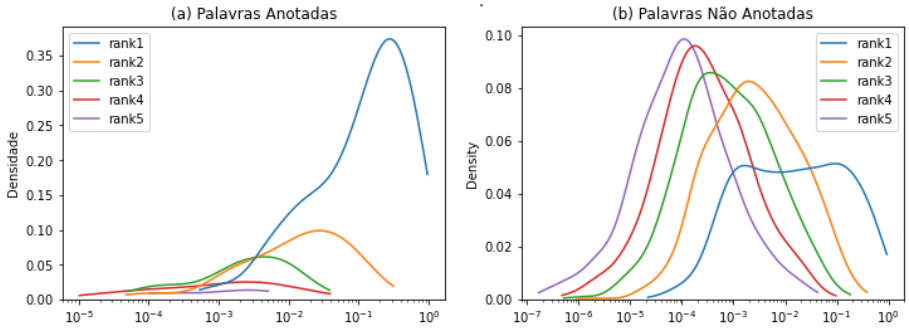


Figura 5. Top 5 palavras relevantes anotadas e não anotadas por rank.

$67,51\%$. Isso indica a possibilidade de que apenas palavras cujo $R > 10^{-2}$ sejam realmente relevantes, tal que as demais possam ser consideradas desnecessárias para o humano entender a classificação da palavra alvo. De forma análoga, na Figura $4(b)$, palavras que não estavam no conjunto anotado cujo valores de $R > 10^{-2}$ representam apenas $18,25\%$ do total, o que corrobora com a análise da Figura $4(a)$, podendo-se considerar as demais como desnecessárias na dialética humana.

Para compreender a diferença entre o rank ordenado por Oclusão e o entendimento humano da Relevância, os valores utilizados na Figura $4(a, b)$ foram separados por posição no rank, ainda em escala logarítmica, e apresentados na Figura $5(a, b)$. As curvas dessa figura indicam as distribuições por densidade de palavras por ordem de grandeza em cada rank dos Top 5 melhores ranqueados. Na Figura $5(a)$ estão apenas as palavras que foram anotadas como relevantes e na Figura $5(b)$ seu complemento.

Verifica-se, na Figura $5(a)$, que a curva do rank 1 apresenta-se bem mais acentuada que as demais e seu pico nos valores de ordem de grandeza de 10^{-1}. Isso indica que a Oclusão tem uma precisão muito alta no rank 1, o que é corroborado pelo achatamento dessa curva na Figura $5(b)$. Os demais ranks possuem curvas bem menos inclinadas na Figura $5(a)$, diminuindo conforme aumenta a posição no rank, mostrando que quanto menor o rank, menor a chance de um humano considerar a palavra relevante. A Figura $5(b)$ evidencia mais essa tendência de erro a medida em que o valor de R e a posição da palavra no rank diminuem.

Como a Oclusão não considera a própria palavra no ranqueamento, devido a forma do cálculo da relevância ser por meio de sua remoção, no caso de não haver valores de R em ordem de grandeza relevante, é possível inferir que a palavra alvo é autossuficiente em sua classificação. Por exemplo, FIFA (Federação Internacional de Futebol) é sempre

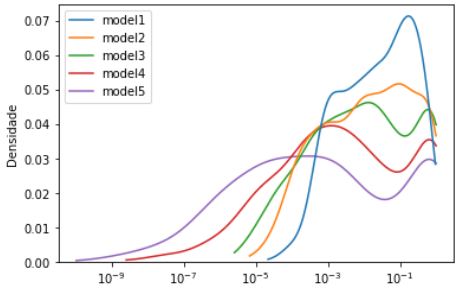


Figura 6. Distribuição dos valores por modelo para o Rank 1.

uma organização, não precisando de palavras da vizinhança para prevê-la. Contudo, essa inferência pode indicar, também, um superajuste do modelo para classificação da palavra alvo, requerendo análises melhores desses casos.

Tais gráficos demonstram que a Oclusão é capaz de distinguir uma palavra relevante de uma não relevante. É importante ressaltar que a medida que a palavra desce no rank, o valor de R também diminui, contudo, esse valor isolado não indica sua posição no rank. Nessa Seção ficou evidente que o entendimento humano de Relevância possui uma maior correlação com a grandeza de R do que o rank da palavra. Dessa forma, a posição do ranking deve ser analisada em conjunto com o valor de relevância.

4.3. Análise do Impacto do Modelo de Aprendizado de Máquina na Explicabilidade

Esta seção apresenta como as alterações do modelo interferem na Explicação. A Figura 6 apresenta a distribuição da palavra mais bem ranqueada em cada frase para os cinco modelos. Como os resultados das palavras da 2ª a 5ª posição no ranking (rank 2 a 5) são similares ao do rank 1, esses resultados não são apresentados neste artigo. Também pode-se perceber que o Modelo 1, com menor treinamento, possui concentração das palavras ranqueadas como mais relevantes com valores de R na ordem de 10^{-1} e, a medida que se aumenta o treinamento dos modelos, esses valores se distribuem em ordens de grandeza menores até que no Modelo 5 elas apresentam-se em maior concentração em 10^{-4}.

Nesses casos, onde os valores de relevância são pequenos, pode-se compreender que o modelo esta superajustado para predição da palavra alvo sendo a mesma autosuficiente para sua classificação. Essa é uma indicação indireta de Relevância da palavra alvo sobre ela mesma, conforme visto na Seção 4.2 e evidenciada pela análise dos modelos.

4.4. Estudos de Casos das Explicações por Oclusão

Esta seção apresenta estudos de casos para verificar se as explicações estão cumprindo seu objetivo. Para equalizar as análises, o Modelo 1, 2 e 3 foram utilizados. Os casos apresentados analisam o erro de predição de uma palavra que deveria ser **B-PER** (nome inicial de pessoa), o erro na explicação de uma palavra **B-PER** que não considera em sua vizinhança uma palavra **I-PER** (nome do meio ou final de uma pessoa), e um exemplo de relevância indireta, no qual nenhuma das palavras apresenta relevância significativa indicando que a mais relevante é a palavra alvo.

Erro de Predição. Na predição da sentença *"there were apparently no adults at the party as the father of the family who lived in the house was out of town and the mother died more than a year ago , black said ."*, a palavra alvo *black* deveria ter rótulo **B-PER**, mas

Tabela 3. Top 5 Relevâncias por Modelo - Estudo de Caso 1.

Rank	Relevância 1		Relevância 2		Relevância 3	
0	black	(5.26E-02)	black	(1.28E-01)	black	(3.14E-01)
1	mother	(9.24E-04)	,	(2.26E-04)	mother	(2.18E-04)
2	ago	(3.10E-04)	ago	(1.42E-04)	,	(1.68E-04)
3	said	(4.21E-05)	mother	(1.11E-04)	house	(1.00E-04)
4	house	(2.78E-05)	.	(7.13E-05)	ago	(7.87E-05)

Tabela 4. Top 5 Relevâncias por Modelo - Estudo de Caso 2.

Rank	Relevância 1		Relevância 2		Relevância 3	
0	tom	(9.94E-01)	tom	(9.99E-01)	tom	(1.00E+00)
1	1.	(3.88E-01)	1.	(3.35E-01)	pukstys	(4.52E-01)
2	pukstys	(1.26E-01)	pukstys	(2.82E-01)	1.	(7.82E-02)
3	(	(7.71E-05)	86.82	(4.78E-05)	86.82	(1.11E-05)
4	86.82	(6.50E-05)	(	(2.41E-05)	u.s.	(1.40E-06)

Tabela 5. Top 5 Relevâncias por Modelo - Estudo de Caso 3.

Rank	Relevância 1		Relevância 2		Relevância 3	
0	mills	(6.78E-01)	mills	(6.60E-01)	mills	(4.41E-01)
1	,	(2.10E-03)	,	(1.95E-03)	lester	(7.69E-04)
2	7:13	(1.55E-03)	pronounced	(8.85E-04)	pronounced	(3.75E-04)
3	father	(1.15E-03)	looked	(8.38E-04)	,	(2.40E-04)
4	1213	(2.21E-04)	father	(1.90E-04)	father	(1.24E-04)

foi predita como **outro** pelos três modelos estudados (Modelo 1, Modelo 2 e Modelo 3). A Tabela 3 apresenta a ordenação das Relevância desse caso. Nas explicações dadas aos três modelos, a palavra mais relevante possui Relevância $R < 4 * 10^{-1}$. Além disso, o primeiro valor da tabela indica o valor absoluto da predição y da palavra alvo, e este também é menor que $4 * 10^{-1}$, o que implica na probabilidade da predição ser menor que 40%. Isto indica que quando a palavra alvo é predita erroneamente com valor de y muito baixo, a explicação para essa palavra também fica comprometida.

Erro de Explicação. Na predição da sentença *"1. tom pukstys (u.s.) 86.82"*, a palavra alvo é *tom*, cujo rótulo é **B-PER**, foi predita corretamente com valor de $y > 9 * 10^{-1}$ pelos três modelos. Na Tabela 4, observa-se que o valor da Relevância R para ambas as palavras ficou na ordem de 10^{-1}, o que indica que foram relevantes na predição da palavra alvo. Contudo, apenas a palavra *pukstys* estava anotada, provavelmente, por ser o sobrenome do *tom* (*pukstys* é um **I-PER**), o que induz o humano à considerá-la como relevante.

Relevância Indireta. Na predição da sentença *"as glenn lawhon , a rural florida minister who is the victim 's father , looked on , mills was pronounced dead at 7:13 a.m. est (1213 gmt) for the murder of lester lawhon ."* a palavra alvo *mills*, cuja rótulo é **B-PER**, foi predita corretamente com valor de $y > 6 * 10^{-1}$ pelos três modelos estudados. Entretanto, não existem palavras anotadas tal que os voluntários as considerassem que alguma outra palavra seria relevante para a classificação de *mills*. Na Tabela 5 pode-se verificar que o valor da Relevância R abaixo de 10^{-2}, o que indica que **mills** foi suficiente para se rotular como **B-PER**. Esse é um dos casos de Explicação indireta que depende da insuficiência de valores de Relevância das demais palavras.

5. Conclusão

Neste trabalho, analisou-se a capacidade da Explicabilidade por Oclusão de explicar resultados da classificação de palavras em tarefas de NER. Para verificar se a Oclusão ofereceu explicações compatíveis com o entendimento humano, foi criado um banco de dados anotado com palavras consideradas relevantes nas frases e esse foi comparado com as explicações por Oclusão. Em uma primeira análise, o rank de Relevância gerado por Oclusão apresentou média da precisão de $65,52\%$ para a palavra mais bem ranqueada seguindo-se para valores que indicam, em média, que pelo menos uma palavra nos top n ranks também foram consideradas relevantes. Isso significa que a explicação por Oclusão é capaz de indicar, em média, pelo menos uma palavra que é Relevante compatível com a compreensão humana. Em seguida, foram analisadas apenas as ocorrências em que a Oclusão concordava, ou não concordava, com a Relevância indicada por humanos. Essa análise revelou, por exemplo, que se não existirem palavras no rank com valor de R acima do valor de corte, a própria palavra alvo pode ser responsável por sua classificação. Finalmente, foi analisado o que ocorre quando uma mesma instância é submetida a modelos com pesos diferentes. Os resultados evidenciaram, que quanto mais se treinou o modelo, mais bem ajustadas são as predições da palavra alvo no sentido da mesma ficar cada vez mais autossuficiente para sua predição. Portanto, a Oclusão é uma técnica capaz de gerar explicações compreensíveis para a classificação de uma palavra.

Limitações e Trabalhos Futuros. A Oclusão pode ser refinada para abranger mais de uma palavra para o cálculo de relevância. Outra análise importante é o caso da palavra alvo depender apenas de si mesma para ser predita. Essa conclusão é obtida de maneira indireta, ou seja, pelo cálculo da Relevância das demais palavras.

Agradecimentos. Este trabalho foi financiado pela Blip (Sucupira S/A).

Referências

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Harbecke, D. (2021). Explaining natural language processing classifiers with occlusion and language modeling. *arXiv preprint arXiv:2101.11889*.

Hu, J. (2018). Explainable deep learning for natural language processing.

Liu, X., Chen, H., and Xia, W. (2022). Overview of named entity recognition. *Journal of Contemporary Educational Research*, 6(5):65–68.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Russell-Rose, T., Stevenson, M., and Whitehead, M. (2002). The reuters corpus volume 1-from yesterday's news to tomorrow's language resources.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Schweter, S. and Akbik, A. (2020). Flert: Document-level features for named entity recognition. *arXiv preprint arXiv:2011.06993*.

Siddharthan, A. (2002). Christopher d. manning and hinrich schutze. foundations of statistical natural language processing. mit press, 2000. isbn 0-262-13360-1. 620 pp. $64.95/£44.95(cloth)$. *Natural Language Engineering*, 8(1):91–92.

Vajjala, S. and Balasubramaniam, R. (2022). What do we really know about state of the art ner? *arXiv preprint arXiv:2205.00034*.

Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Wallace, E., Feng, S., and Boyd-Graber, J. (2018). Interpreting neural networks with nearest neighbors. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 136–144, Brussels, Belgium. Association for Computational Linguistics.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., and Tu, K. (2020). Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.