Alexandre Augusto Aguiar Gomes

Uso de Oclusão para Explicabilidade em Inteligência Artificial Aplicada em Reconhecimento de Entidades Nomeadas

Belo Horizonte

2023

Alexandre Augusto Aguiar Gomes

Uso de Oclusão para Explicabilidade em Inteligência Artificial Aplicada em Reconhecimento de Entidades Nomeadas

Projeto de Mestrado apresentado ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Centro Federal de Educação Tecnológica de Minas Gerais – CEFET-MG

Modelagem Matemática e Computacional

Sistemas Inteligentes

Orientador: Flávio Luis Cardeal Padua

Coorientador: Daniel Hasan Dalip

Belo Horizonte

2023

Espaço destinado à folha de aprovação

Dedico este trabalho aos professores e colabores do Centro de Formação Tecnológica de Minas Gerais que acreditam que o conhecimento é uma ferramenta que precisa ser constantemente polida e afiada. Faço uma dedicação especial aos professores, colaboradores e parceiros que, além disso, acreditam que a academia pode ser parceira da iniciativa privada para o desenvolvimento de tecnologias que possam fazer do nosso mundo um lugar melhor para todos. Faço, ainda, uma dedicação póstuma ao meu pai que nos deixou e não podendo presenciar este momento.

Agradecimentos

Passamos por anos difíceis mas felizmente conseguimos chegar nesse momento, agradeço, então, aos professores e colabores do Programa de Pós Graduação do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), que mesmo com o advento de uma pandemia, se dispuseram a ajudar e manter o programa; à equipe da Take-Blip que se dispôs a compartilhar conhecimento e disponibilizar bolsa de pesquisa para este projeto; aos meus pais e irmão por me darem o conhecimento base para chegar à esse momento; e à minha amada esposa e filhos, que, mesmo nos momentos mais difíceis, me suportaram e acreditaram em mim.

'Os indivíduos se diferem pela habilidade de entender ideias complexas, de se adaptarem com eficácia ao ambiente, de aprenderem com a experiência, de se engajarem nas várias formas do raciocínio e de superarem obstáculos mediante o pensamento.'

(Associação Americana de Psicologia, 1995)

Resumo

Técnicas de Explicabilidade de Inteligência Artificial (XAI, do inglês *Explainable Artificial Intelegence*) são métodos que auxiliam o usuário a entender os motivos pelos quais o modelo de Aprendizado de Máquina (ML, do inglês, *Machine Learning*) realizou uma determinada predição. 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 e também na determinação de quais palavras são relevantes para efetuar uma determinada predição automática, por exemplo, nas tarefas de identificação automática de classes gramaticais ou reconhecimento de entidades nomeadas (NER, do inglês, *Named Entity Recognition*). Esse trabalho tem como objetivo estudar a aplicabilidade da técnica de XAI de Oclusão em NER. A Oclusão, ao ser aplicada em NER, é uma técnica que mensura a relevância de uma determinada palavra na predição por meio da análise do impacto em sua remoção. Essa técnica, além de trazer um entendimento do processo de predição, pode permitir que sejam realizadas intervenções no ajuste dos hiperparâmetros do modelo ou no preprocessamento da base de dados. Para realizar esse trabalho, utilizou-se uma LSTM bidirecional e o dataset CoNLL 2003, amplamente utilizado na literatura. Logo após, foi investigado a precisão dessa técnica de explicabilidade analisando a capacidade da mesma em gerar respostas similares às esperadas por humanos. Para realizar essa análise, foi utilizado a anotação manual 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. Dessa forma, tanto a análise realizada quanto a base de dados de referência são contribuições do presente trabalho.

Palavras-chave:Aprendizado de Máquina. Explicabilidade em Inteligência Artificial. Reconhecimento de Entidades Nomeadas. Oclusão.

Abstract

Explainability in Artificial Intelligence Techniques (XAI) are methods that help the user to understand the reasons why the Machine Learning model (ML) made a prediction. They have been applied to a wide range of problems, from Computer Vision to Natural Language Processing (NLP). In NLP, such techniques can allow the user to identify which words are important for understanding a sentence and also to determine which words are relevant to perform a certain automatic prediction, for example, in tasks of automatic identification of parts of speech word or Named Entities Recognition (NER). This work aims to study the applicability of the XAI Occlusion technique in NER. Occlusion, when applied in NER, is a technique that measures the relevance of a given word prediction by analyzing the impact of its removal. This technique, in addition to bringing an understanding of the prediction process, can allow interventions to be carried out in the adjustment of the hyperparameters of the model or in the preprocessing of the database. To carry out this work, a bidirectional LSTM and the CoNLL 2003 dataset, widely used in the literature, were used. Soon after, the accuracy of this explainability technique was investigated by analyzing its ability to generate responses similar to those expected by humans. To carry out this analysis, manual annotation of a set of 849 sentences was used, thus creating a reference database. As far as we know, there are no analyzes in the literature of the use of Occlusion in NER. Thus, both the analysis carried out and the reference database are contributions of this work.

Keywords: Machine Learning, Explainability Artificial Intelligence, Named Entities Recognition, Occlusion.

Lista de ilustrações

Figura 1 – Exemplo de Classificação de uma **Localidade** por NER. 25
Figura 2 – Exemplo de Classificação de uma Entidade que Possui Ambiguidade. . 26
Figura 3 – Exemplo da Ordenação por Relevância das Palavras da Entrada 1 da Figura 2. 27
Figura 4 – Exemplo da Entidade *São Paulo* sendo Generalizada de Três Formas Diferentes. 32
Figura 5 – Detalhe de uma Camada em Arquitetura LSTM. 33
Figura 6 – Exemplo de uma Arquitetura BiLSTM. 34
Figura 7 – Exemplo Ilustrativo da Perturbação da Vizinhança por Oclusão. 37
Figura 8 – Exemplo Ilustrativo de Ordenação de Relevâncias Obtidos por Oclusão. 37
Figura 9 – Fluxo da Metodologia 41
Figura 10 – Rotulador *Web* Desenvolvido no Escopo deste Trabalho. 45
Figura 11 – Precisão 50
Figura 12 – Distribuição das Top5 Palavras Relevantes Anotadas por Ordem de Grandeza. 51
Figura 13 – Densidade das Top5 Palavras Relevantes Anotadas e Não Anotadas por Rank. 52
Figura 14 – Distribuição dos Valores por Modelo - Rank 1. 53
Figura 15 – Distribuição dos Valores por Modelo - Ranks 2, 3, 4, 5. 54
Figura 16 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 5 Épocas. . 73
Figura 17 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 10 Épocas. 74
Figura 18 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 20 Épocas. 75
Figura 19 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 50 Épocas. 76
Figura 20 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 100 Épocas. 77
Figura 21 – Distribuição dos Valores do Modelo 1 - Ranks 2, 3, 4, 5. 83
Figura 22 – Distribuição dos Valores do Modelo 2 - Ranks 2, 3, 4, 5. 84
Figura 23 – Distribuição dos Valores do Modelo 3 - Ranks 2, 3, 4, 5. 84
Figura 24 – Distribuição dos Valores do Modelo 4 - Ranks 2, 3, 4, 5. 85
Figura 25 – Distribuição dos Valores do Modelo 5 - Ranks 2, 3, 4, 5. 85

Lista de tabelas

Tabela 1 – Abrangência das Explicações de uma XAI, de Acordo com Guidotti et al. (2018). 35
Tabela 2 – Classificação para o Propósito da XAI, Segundo Guidotti et al. (2018). 35
Tabela 3 – Classificação das Técnicas de Coleta da Informação de XAIs, de Acordo com Guidotti et al. (2018). 36
Tabela 4 – Características de uma Explicação, Segundo Ribeiro, Singh e Guestrin (2016). 37
Tabela 5 – Distribuição do CoNLL 2003 inglês. 42
Tabela 6 – Modelos Utilizados para Avaliar as Explicações. 43
Tabela 7 – Distribuição da Amostra do CoNLL 2003 Anotada Manualmente. . . . 45
Tabela 8 – Modelos Utilizados para Avaliar as Explicações. 53
Tabela 9 – Predições por Modelo -Estudo de Caso 1. 55
Tabela 10 – Relevância Ordenadas por Modelo -Estudo de Caso 1. 56
Tabela 11 – Predições por Modelo -Estudo de Caso 2. 57
Tabela 12 – Relevância Ordenadas por Modelo -Estudo de Caso 2. 57
Tabela 13 – Predições por Modelo -Estudo de Caso 3. 58
Tabela 14 – Relevância Ordenadas por Modelo -Estudo de Caso 3. 59
Tabela 15 – Modelos Utilizados para Avaliar as Explicações. 71
Tabela 16 – Valores das Precisões nas k_esimas Posições Utilizados no Gráfico de Precisão. 79
Tabela 17 – Quantidade de Palavras Anotadas e Não Anotadas por Posição no Rank. 81
Tabela 18 – Porcentagem de Palavras Anotadas e Não Anotadas por Posição no Rank. 81

Lista de abreviaturas e siglas

biLSTM	do inglês *Bidirecional Long Short - Term Memory*, Memória de Termo - Longo Curto Bidirecional;
IA	Inteligência Artificial;
LRP	do inglês *Layer-Wise Relevance Propagation*, Propagação da Relevâcia entre Camadas;
LSTM	do inglês *Long Short - Term Memory*, Memória de Termo - Longo Curto;
ML	do inglês *Machine Learning*, Aprendizado de Máquinas;
NER	do inglês *Named Entity Recognition*, Reconhecimento de Entidades Nomeadas;
NLP	do inglês *Natural Language Processing*, Processamente de Linguagem Natural;
XAI	do inglês *Explainability in Artificial Inteligence*, Explicabilidade em Inteligência Artificial.

Lista de símbolos

$\overline{x}$	Barra sobreposta indicando média da variável que a acompanha;
Σ	Letra grega Sigma usada para indicar uma sequência de somas.

Sumário

1 INTRODUÇÃO . . . 25
1.1 Objetivos . . . 28
1.2 Contribuições . . . 28
1.3 Organização do Texto . . . 28

2 FUNDAMENTAÇÃO TEÓRICA . . . 31
2.1 Sumarização e Reconhecimento de Entidades Nomeadas . . . 31
2.2 Aprendizado de Máquina e arquitetura biLSTM . . . 33
2.3 Explicabilidade em Inteligência Artificial (XAI) . . . 34
2.4 Explicabilidade por Oclusão . . . 36

3 TRABALHOS RELACIONADOS . . . 39

4 METODOLOGIA . . . 41
4.1 Base de Dados de NER . . . 42
4.2 Modelo de Aprendizado de Máquina (biLSTM) . . . 42
4.3 Implementação da Oclusão . . . 44
4.4 Análise dos Resultados . . . 44
4.4.1 Amostra Anotada Manualmente com Palavras Relevantes . . . 44
4.4.2 Métrica de Avaliação . . . 46

5 RESULTADOS . . . 49
5.1 Análise Precisão da Explicação por Oclusão considerando a Explicação Humana . . . 49
5.2 Análise dos Valores de Relevância Estimados pela Oclusão . . . 50
5.3 Análise do Impacto do Modelo de Aprendizado de Máquina na Explicabilidade . . . 52
5.4 Estudos de Casos das Explicações por Oclusão . . . 54
5.4.1 Erro de Predição . . . 55
5.4.2 Erro de Explicação . . . 57
5.4.3 Relevância Indireta . . . 57

6 CONCLUSÃO . . . 61

REFERÊNCIAS . . . 65

APÊNDICES 69

APÊNDICE A – LOSS E ACURÁCIA DOS MODELOS UTILIZADOS 71

APÊNDICE B – MATRIZES DE CONFUSÃO DOS MODELOS UTILIZADOS 73

APÊNDICE C – VALORES DAS PRECISÕES NAS K POSIÇÕES 79

APÊNDICE D – PALAVRAS ANOTADAS E NÃO ANOTADAS POR POSIÇÃO NO RANK 81

APÊNDICE E – GRÁFICOS DA PALAVRAS RELEVANTES POR RANK SEPARADOS POR MODELO. 83

1 Introdução

Introduzida por Lent, Fisher e Mancuso (2004), a Explicabilidade em Inteligência Artificial (XAI, do inglês, *Explainable Artificial Intelegence*) é uma ferramenta utilizada por usuários de Redes Neurais que o auxilia a entender os motivos pelos quais o modelo de Aprendizado de Máquina (ML, do inglês, *Machine Learning*) realizou uma determinada predição. 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 automática, por exemplo, nas tarefas de identificação automática de classes gramaticais (PoS taggig, do inglês, *Part of Speech tagging*) ou Reconhecimento de Entidades Nomeadas (NER, do inglês, *Named Entity Recognition*).

O objetivo de NER aplicada em NLP é identificar palavras ou expressões de forma que seja mais fácil entender e processá-los. Por ser uma técnica de ML o interesse por sua aplicação cresceu muito dada a grande quantidade de informação disponível na *internet*. Esta técnica foi inicialmente aplicada na classificação de nomes próprios que indicavam pessoa, localidade ou organização, mas foi rapidamente estendida para a compreensão e classificação de termos técnicos médicos e científicos, se estendendo em seguida para diversas outras áreas. Na Figura 1 o nome composto *São Paulo*, por exemplo, foi classificado como **Localidade** por um modelo de ML para NER.

Tal técnica é importante pois a classificação dessas entidades nem sempre é direta, sendo frequentemente analisado o contexto da frase. No exemplo da Figura 2 são apresentadas outras duas possibilidades de classificação para *São Paulo*. Além de **Localidade**, como apresentado na Figura 1, ele pode receber a classificação de **Organização** (Prefeitura de São Paulo) ou **Pessoa** (Santo católico), podendo-se dizer que essa entidade apresenta uma ambiguidade.

A ambiguidade de palavras na utilização de NER é um dos principais desafios no qual o estado da arte para esse problema seria utilizando. Nesses casos, um modelo de Inteligência Artificial seria capaz de interpretar uma frase (instância) e predizer qual seria

Figura 1 – Exemplo de Classificação de uma **Localidade** por NER.

a melhor classificação para uma entidade ambígua dada aquela instância (TORFI et al., 2020). Geralmente, utilizam-se Redes Neurais Artificiais e arquiteturas do tipo biLSTM (do inglês (*Bidirectional Long Short - Terms Memory*) (QI, 2018) para tais tarefas. Tais modelos possuem uma precisão consideravelmente alta (acima de $\sim 90\%$ (VAJJALA; BALASUBRAMANIAM, 2022)).

No entanto, é sempre importante que usuários/desenvolvedores entendam o porquê de uma determinada classificação. A explicabilidade pode prover ao usuário tais explicações permitindo uma melhor tomada de decisão e cumprir, inclusive, legislações de Proteção de Dados (GOODMAN; FLAXMAN, 2017; MALGIERI; COMANDÉ, 2017; WACHTER; MITTELSTADT; FLORIDI, 2017). Além disso, por meio delas, pode-se perceber tendências discriminatórias (PEDRESHI; RUGGIERI; TURINI, 2008) e mitigar problemas causados por falhas de predição (GUIDOTTI et al., 2018).

Uma das possibilidades de se explicar o resultado da predição de um modelo de ML é analisar a interação da palavra estudada com as entidades presentes na instância. Especificamente em NER, pode-se analisar o impacto que as demais palavras da frase tem na nomeação da entidade a ser classificada.

As técnicas de XAI em NER surgiram dessa necessidade e consistem em indicar quais são palavras que mais colaboraram para que determinada entidade seja classificada, apresentando representações eficientes da construção dos resultados da predição de modelos de ML, considerados caixas pretas (característica de um modelo no qual o usuário conhece apenas a entrada e a saída do modelo, mas não sabe como o modelo fez para chegar na saída a partir daquela entrada).

Conforme mostrado no exemplo da Figura 2, o termo *São Paulo* poderia receber uma das três classificações, **Pessoa**, **Localidade** ou **Organização**. A Figura 3 apresenta o ranking que poderia ter sido gerado por meio de uma técnica de Explicabilidade auxiliando, assim, a compreender melhor porque na Entrada 1 desse exemplo a entidade foi classificada como **Localidade** em detrimento aos demais possíveis rótulos.

Neste exemplo, a explicação apontou a palavra '*para*' como sendo a mais relevante para a classificação. Outras aplicações em NLP poderiam considerá-la uma *Stop Word* (palavra dispensável para o entendimento da informação), contudo, para NER, algumas

Figura 2 – Exemplo de Classificação de uma Entidade que Possui Ambiguidade.

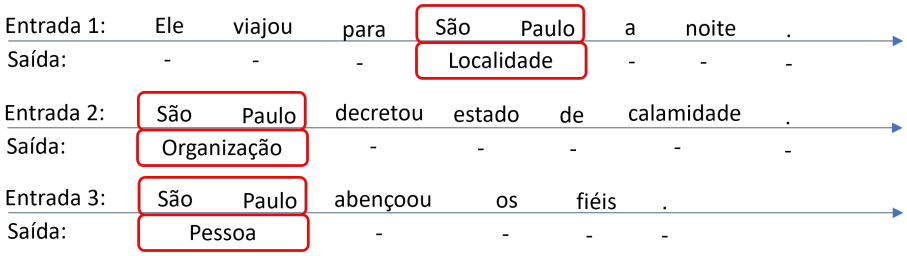


Figura 3 – Exemplo da Ordenação por Relevância das Palavras da Entrada 1 da Figura 2.

Palavra	Relevância
para	$r_3 = 0{,}20$
Paulo	$r_4 = 0{,}10$
viajou	$r_2 = 0{,}09$
a	$r_5 = 0{,}09$
noite	$r_6 = 0{,}01$
Ele	$r_1 = 0{,}01$
.	$r_7 = 0{,}00$

dessas palavras trazem informações de conectivos importantes, tal como é a palavra '*para*', que traz sentido de **Localidade** quando acompanhada do verbo '*viajar*', pois "quem viaja para, viaja para algum lugar", excluindo as possibilidades de uma **Pessoa** ou **Organização**, como exemplificado na Figura 2.

Além de conceder ao usuário uma justificativa do motivo de uma classificação ser realizada, uma boa explicação poderia prover auxilio caso do modelo falhe em sua predição, uma vez que ela é capaz de apontar as palavras que tem maior relevância para classificação da entidade, permitindo identificar, por outras formas, o motivo de uma possível interação inesperada.

As técnicas de Explicabilidade podem variar de acordo com seu domínio (dependência ou não do modelo de ML) ou abordagem na elaboração das explicações (engenharia reversa ou perturbação da vizinhança). A proposta deste trabalho é apresentar uma implementação da técnica de Oclusão aplicada para NER. Essa técnica remove uma palavra da instância e mensura o impacto de sua remoção, por meio de sua Relevância R, na predição de uma determinada palavra. Além disso, ela pode trazer um entendimento do processo de predição, permitindo que sejam realizadas intervenções no ajuste dos hiperparâmetros do modelo ou no preprocessamento da base de dados.

Para fazer a classificação automática das entidades, este trabalho utilizou-se uma LSTM bidirecional e o dataset CoNLL 2003, amplamente utilizados na literatura. Logo após, foi investigado 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*, 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. Dessa forma, tanto a análise realizada quanto a base de dados de referência são contribuições do presente trabalho, assim como a aplicação *Web* utilizada na anotação dos dados.

1.1 Objetivos

O objetivo deste trabalho é implementar e analisar a técnica de Oclusão para Explicabilidade em Inteligência Artificial aplicada no Reconhecimento de Entidades Nomeadas, no qual foi utilizada uma LSTM bidirecional para a classificação dessas entidades. Será utilizado um conjunto de dados anotados manualmente, criado no âmbito deste projeto, com o objetivo de verificar a precisão desta implementação em gerar explicações similares a humanos por meio de métricas de avaliação.

Essas análises permitirão responder: 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 estudo futuras?

Trabalhos futuros podem realizar novas abordagens na utilização da Oclusão em NLP, assim como a comparação de outras técnicas de Explicabilidade à ela. Além disso, a metodologia dessas análises podem direcionar futuras avaliações de Explicações uma vez que foram desenvolvidas no âmbito desse trabalho.

1.2 Contribuições

Como contribuições, o presente trabalho planeja de disponibilizar, além da implementação da Oclusão aplicada em NER, o banco de dados anotado manualmente, como referência, e as análises realizadas para estudar esta XAI, assim como uma aplicação *Web* que permitiu a anotação dos dados de forma mais célere.

1.3 Organização do Texto

Este trabalho esta distribuído em:

1. Introdução: Apresentação do problema de NER e da Explicabilidade, motivações e desafios, objetivos e contribuições deste trabalho para a área;

2. Fundamentação Teórica: Descrição dos problemas de NER, ferramentas de ML que possibilitam trata-los e uma descrição do processo de Explicabilidade (XAI), incluindo a técnica utilizada nesse trabalho;

3. Trabalhos Relacionados: Um apanhado de trabalhos na área de NLP e XAI atuais que podem auxiliar no estudo deste trabalho;

4. Metodologia: Uma descrição do trabalho técnico apresentado neste texto, desde escolha e descrição do dataset e do modelo de ML, caracterização da Oclusão, a criação de um banco de dados de referência e as estratégias para analisar a XAI;

5. Resultados: Apresentação e análise dos resultados quanto a Precisão da Oclusão em gerar respostas similares às dos humanos, uma análise do valores de Relevâncias os quais humanos consideram uma palavra Relevante e os impactos da qualidade do modelo de ML nos resultados;

6. Conclusões: Serão apresentadas as considerações com base nos resultados alcançados, desafios e futuros trabalhos com base nos resultados;

7. Referências: Trabalhos que serviram de base de conhecimento para o desenvolvimento deste.

2 Fundamentação Teórica

O objetivo deste trabalho é estudar a aplicabilidade da Oclusão em NER. Guidotti et al. (2018) define diversas abordagens nas quais uma XAI pode ser estudadas, dentre elas, sua capacidade de interpretabilidade/fidelidade e a forma como se obtém a explicação. Esse cálculo pode ser realizado por de engenharia reversa, a partir das predições realizadas, ou perturbações de instâncias, sendo essa, a estratégia utilizada na Oclusão. Essas, e outras abordagens, serão melhor detalhadas na Seção 2.3.

Neste capítulo, também serão apresentados os principais conceitos para compreensão deste trabalho, são eles: (*i*) sumarização e Reconhecimento de Entidades Nomeadas 2.1; (*ii*) Aprendizado de Máquina e arquitetura biLSTM para NER 2.2; (*iii*) as características e abordagens de uma XAI (Seção 2.3); (*iv*) e os conceitos no desenvolvimento da Oclusão para seu uso em NER (Seção 2.4).

2.1 Sumarização e Reconhecimento de Entidades Nomeadas

Para melhor compreender os objetivos do NER, faz-se necessário um aprofundamento nas técnicas de sumarização aplicadas à textos. Sumarização refere-se à tentativa de condensar um texto em informações suficientes para o sua compreensão, que pode variar desde sumários (JONES, 1999) indexados à resenhas ou resumos (PAICE; JONES, 1993). Hutchins (1987) separou as informações sumarizadas em dois grandes grupos, enquanto as macro estruturas destinam-se a apresentar relações entre os blocos de sentenças e a organização textual, as micro estruturas consideram as interações entre as sentenças do texto. Hutchins também sugere quatro macro regras que podem ser utilizadas para generalizar ou condensar um texto.

1. Delição: exclusão de informações consideradas irrelevantes ao entendimento da sentença;

2. Generalização: substituição de palavras ou expressões muito específicas por ideias mais abstratas que os representem;

3. Construção: sintetização de um conjunto de informações, semelhante à generalização, mas de forma mais abrangente;

4. Composição delição-construção: delição de sentenças que expressam pré condições ou motivações para a descrição de ações podendo, ou não, reconstruir a ideia principal.

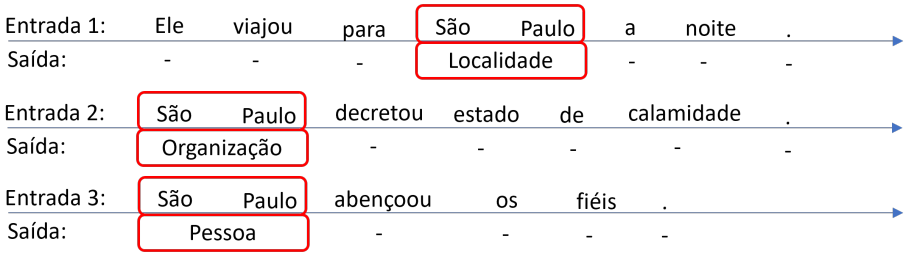
A quantidade de informações disponíveis com a popularização da internet fez crescer a necessidade da estruturação de textos. A medida que a Recuperação de Informações aumentou o volume de dados extraídos das diversas nuvens, a estruturação manual desses textos por meio da sumarização tornou-se inviável (GRISHMAN; SUNDHEIM, 1996). Dessa forma, a sumarização automática utilizando NLP é indispensável, tal que NER tornou-se uma das principais ferramentas para a generalização (HUTCHINS, 1987) de textos permitindo maior celeridade em seu processamento.

Em NER, textos em linguagem natural são submetidos a uma generalização (HUTCHINS, 1987) sendo reduzidos e estruturados de forma que possam ser usados posteriormente por ferramentas de busca de forma rápida e eficiente. Essa proposta se assemelha ao conceito de Designações Rígidas de Kripke (KRIPKE; DAVIDSON; HARMAN, 1972), onde as entidades são diretamente mapeadas às suas generalizações. Entretanto, uma generalização adequada depende do uso correto de regras gramaticais, de habilidades de inferência lógica, vasto conhecimento do vocabulário, além do próprio conhecimento linguístico (SAMPSON, 1987).

A utilização dos conceitos linguísticos apontada por Sampson (1987) pode fazer com que uma entidade perca seu caráter rígido (em que cada entidade é mapeada diretamente para sua generalização) o que amplia as interpretações de contexto (ou seja, uma mesma entidade pode assumir mais de uma generalização dependendo do contexto onde foi usada), como pôde-se ver no Exemplo da Figura 4 em que a entidade *São Paulo* pode assumir três generalizações diferentes, dependendo da frase.

Em linguística, a capacidade de uma palavra assumir mais de um significado é chamada de ambiguidade. Para ML, a ambiguidade é tida como uma incapacidade de se mapear diretamente uma entidade ao seu rótulo, sendo necessária a interpretação de sua vizinhança para se obter uma predição que faça sentido. Tal que essa é a melhor forma de alcançar um resultado linguisticamente correto.

Figura 4 – Exemplo da Entidade *São Paulo* sendo Generalizada de Três Formas Diferentes.



2.2 Aprendizado de Máquina e arquitetura biLSTM

O Aprendizado de Máquina (ML, do inglês *Machine Learning*) aplicado a NER pode de emular boa parte da experiência humana no que consiste o conhecimento linguístico, vocabulário e até regras gramaticais por meio de treinamento por hábito (treinamento baseado em banco de dados de casos). No entanto, existe um desafio para o computador em conseguir simular a inferência lógica humana aplicada às sentenças cuja entidades apresentem ambiguidade.

Existem, na literatura, diversas formas de tratar este problema usando variadas arquiteturas de ML, contudo, neste trabalho será utilizada a LSTM bidirecional, implementação baseada no trabalho de Hu (2018), que aborda a utilização da técnica de LRP para explicação das predições da LSTM.

Analisar a relação entre as palavras de uma sentença pode ser uma solução para o problema da ambiguidade. A arquitetura LSTM, proposta por Hochreiter e Schmidhuber (1997), é uma das ferramenta utilizadas para interpretar a dependência da vizinhança (QI, 2018). Essa arquitetura utiliza uma complexa rede de entradas e saídas com *Gates* que são responsáveis por avaliar os pesos das palavras anteriores e definir se ela é relevante para a rotulação. Esses pesos são armazenados em memória em caso positivo ou descartados caso contrário.

Na Figura 5, pode-se ver como os pesos C e a saída h da camada anterior, percorrem a camada atual passando por operadores de exclusão (ou esquecimento) e agregadores. Dessa forma, os pesos são recalculados com novos valores de entrada x que atualizam os valores de C e h, em memória, resultando na predição de saída h.

Essa arquitetura permite que apenas entidades relacionadas interajam umas com as outras evitando que ruídos interfiram no cálculo dos pesos e da predição. A característica da

Figura 5 – Detalhe de uma Camada em Arquitetura LSTM.

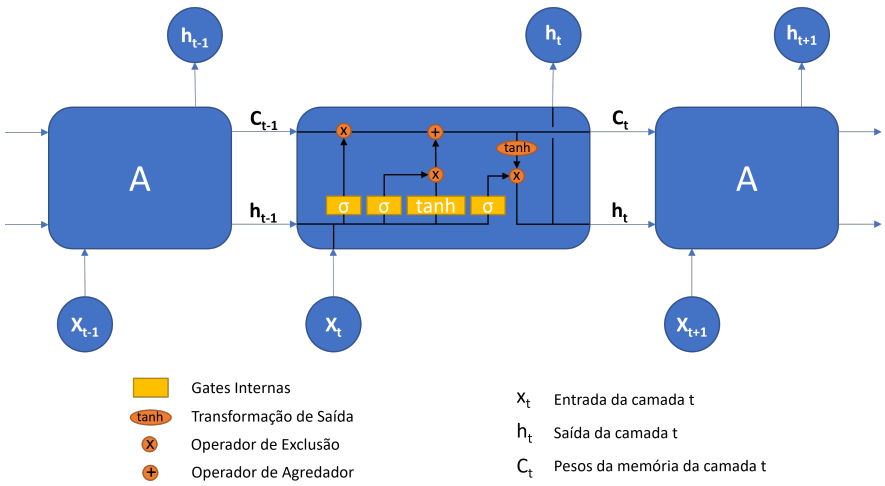
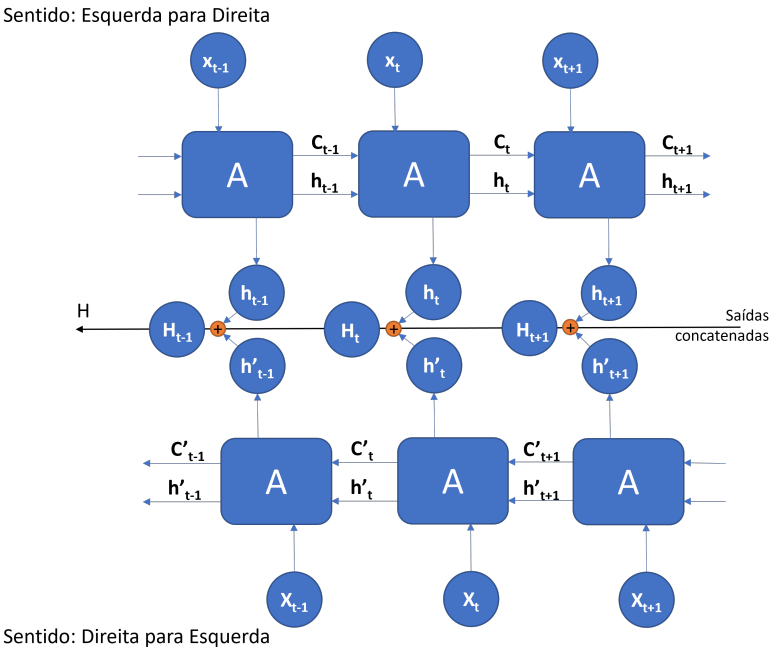


Figura 6 – Exemplo de uma Arquitetura BiLSTM.



LSTM em ponderar as dependências da vizinhança da palavra alvo possibilita uma predição de forma semelhante à inferência lógica humana, ou seja, analisar as palavras próximas e entender a qual contexto ela pertence. Assim como pode-se ver no exemplo da Entrada 1 na Figura 4, na qual deduz-se que *São Paulo* é uma **Localidade** porque as palavras vizinhas à esquerda são 'viajou' e 'para' consecutivamente, exigindo um substantivo de localidade em sequência.

A Figura 6 apresenta uma arquitetura bidirecional para a LSTM (GRAVES; SCHMIDHUBER, 2005), a qual permite que o modelo possa analisar as dependências tanto das palavras à esquerda, no sentido tradicional da leitura, quanto à direita da palavra alvo. As saídas em cada sentido são agregadas em um único valor de saída H.

Com essa arquitetura, consegue-se interpretar, também, as vizinhanças após a palavra alvo, como seria o caso das Entrada 2 e 3 da Figura 4, as quais pode-se deduzir que *São Paulo* é uma **Organização** devido a palavra 'decretou', que vem logo em seguida e **Pessoa** devido à 'abençoou', de forma análoga.

2.3 Explicabilidade em Inteligência Artificial (XAI)

Segundo Samek, Wiegand e Müller (2017), a explicação de uma XAI é uma tentativa do ser humano em explicar para outro ser humano como uma máquina aprendeu a predizer a solução de um problema. Essa necessidade surgiu do fato de precisarmos compreender a base de um raciocínio para confiarmos nas informações resultantes desse raciocínio. De maneira geral, os modelos de ML são vistos como caixa preta (GUIDOTTI et al.,

Tabela 1 – Abrangência das Explicações de uma XAI, de Acordo com Guidotti et al. (2018).

Abrangência	Explicação
Global	A explicação tem como objetivo principal explicar a relação entre as entradas e as saídas sem se preocupar, necessariamente, com o que ocorre dentro do modelo.
Local	A explicação busca indicar quais as decisões internas na NN que foram responsáveis pela predição.

2018) evidenciando a necessidade da utilização de uma técnica capaz esclarecer o que ocorre dentro deles. Por esse motivo, uma XAI busca simular o que ocorre com modelo de ML durante a predição permitindo ao usuário utilizá-la conferindo explicações para os resultado ou estudar possibilidades de melhorias.

Complementarmente, para Guidotti et al. (2018), a XAI provê representações para que seja possível compreender propriedades específicas da predição, como o comportamento ao se perturbar um ou mais atributos, ou identificações de pesos ou componentes do modelo responsáveis por determinadas decisões. A Tabela 1 descreve como as XAIs podem variar conforme sua abrangência ou seu propósito em relação ao modelo. Uma abrangência 'Global' significa que a explicação é focada em explicar qual a relação entre as entradas e a saída predita, enquanto as 'locais' são capazes de percorrer o modelo e apontar quais as decisões do modelo foram responsáveis para determinar a solução.

As XAIs podem ser utilizadas não somente para entender a IA, mas com propósitos técnicos na ciência de dados ou no ML. Usada de forma adequada, elas podem correlacionar os dados de entrada do treinamento com as predições obtidas (Explicador de Amostra), explicar as correlações entre os dados dentro do modelo (Explicador de Modelo) ou até mesmo verificar se o modelo esta funcionando como desejado (Inspeção do Modelo) (GUIDOTTI et al., 2018). A Tabela 2 descreve esses propósitos segundo abordagem de Guidotti et al. (2018).

A forma como engenharia de dados explica o processo vai depender da abrangência

Tabela 2 – Classificação para o Propósito da XAI, Segundo Guidotti et al. (2018).

Propósito	Explicação
Explicador de Amostra	Consiste em encontrar uma correlação entre os dados de entrada do modelo com a tomada de decisão de forma local.
Explicador de Modelo	Consiste em explicar de forma global o modelo.
Inspeção do Modelo	Consiste em utilizar de estratégias globais e/ou locais para visualizar aspectos específicos do modelo.

e do propósito da explicação. A Tabela 3 descreve essas técnicas de uma forma genérica, conforme abordagem de Guidotti et al. (2018). No caso da explicação ser essencialmente global, cujo o modelo não é o foco, técnicas de Perturbações das Instâncias de Entrada (ou seja, pequenas variações de uma mesma amostra são inseridas no modelo como teste) podem ser suficientes para obter uma explicação adequada ao propósito, contudo, quando deseja-se detalhes das camadas do modelo, técnicas de Engenharia Reversa podem ser necessárias (GUIDOTTI et al., 2018).

A Tabela 4 descreve com detalhes as características de uma explicação atendendo aos objetivos do usuário. Ela deve ser capaz de interpretar um modelo de modo a reduzir as dúvidas do usuário a cerca de como uma predição foi alcançada e/ou ser minimamente fiel ao modelo, permitindo ao usuário ser capaz de definir em qual ou quais camadas do modelo a predição foi afetada (RIBEIRO; SINGH; GUESTRIN, 2016; GUIDOTTI et al., 2018).

As próximas sub sessões descrevem os conceitos aplicados na Oclusão utilizados nesse trabalho.

2.4 Explicabilidade por Oclusão

Descrita por Robnik-Šikonja e Kononenko (2008), a técnica de Oclusão para Explicabilidade consiste em remover parte da informação de entrada e medir o impacto dessa remoção no valor da predição. Ela se baseia na estratégia da Perturbação da Vizinhança descrita por Guidotti et al. (2018), utilizando o próprio modelo para inspecionar as ativações internas e de abrangência global. No exemplo da Figura 7, a instância 0, sem a Perturbação, apresenta predição y, para a entidade *São*, com o valor de $0,99$ para **Localidade**. Na instância i, após a remoção da palavra 'para', o valor foi de $y = 0,79$ e, na instância k, sem a palavra 'Paulo', o valor de $y = 0,89$.

A diferença do valor da predição y indica que a perturbação teve algum impacto no processo. Quanto maior a variação de y na saída, maior a relevância das palavras removidas

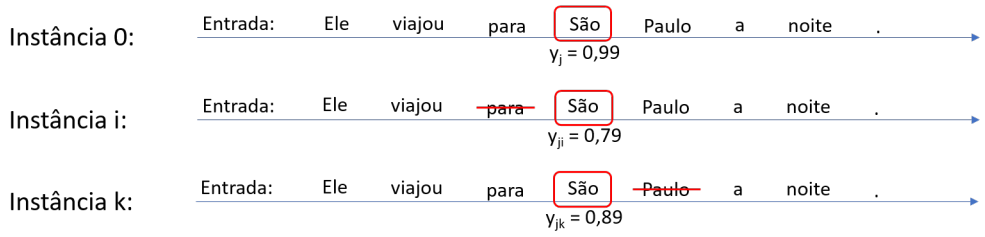
Tabela 3 – Classificação das Técnicas de Coleta da Informação de XAIs, de Acordo com Guidotti et al. (2018).

Técnica	Explicação
Engenharia Reversa	Consiste em utilizar um modelo auxiliar para simular com mais transparência e menos complexidade o modelo original.
Perturbação de Instância	Consiste em utilizar o próprio modelo para se inspecionar suas ativações internas por meio de Perturbações na entrada.

Tabela 4 – Características de uma Explicação, Segundo Ribeiro, Singh e Guestrin (2016).

Característica	Explicação
Interpretabilidade (Global)	É a capacidade da técnica em explicar o resultado da predição para o usuário da ferramenta. Essa característica é muito dependente da compreensão do usuário do problema. Portanto, muito susceptível aos viesses do usuário.
Fidelidade (Local e Global)	É a capacidade da técnica em reproduzir local ou globalmente o resultado de uma predição segundo o modelo. Essa característica esta relacionada apenas às mecânicas do modelo (treino e pesos), portanto, reflete apenas o aprendizado e as relações entre os atributos do modelo.

Figura 7 – Exemplo Ilustrativo da Perturbação da Vizinhança por Oclusão.



no processo de Oclusão. Algumas possibilidades dessa técnica incluem a remoção de janelas de atributos, na intensão de identificar um conjunto atributos que só fazem sentido juntos (como é o caso de pixels de uma imagem) ou um conjunto de palavras que tem significado de uma expressão (LIU, 2019; HEWITT; MANNING, 2019).

A varição dos valores de y obtidos pela técnica de Oclusão, são ordenados e apresentados como Relevâncias das palavras na classificação da entidade alvo, de forma a indicar quais foram mais relevantes para a predição como exemplificado na Figura 8.

Na exemplo da Figura 8, ilustrando uma possível Explicação para o caso da Figura 7, a palavra 'para' teria a maior contribuição na a rotulação de *São* como **Localidade**, enquanto o ponto final, '.', seria a menor.

Figura 8 – Exemplo Ilustrativo de Ordenação de Relevâncias Obtidos por Oclusão.

Palavra	Relevância
para	r_3 = 0,20
Paulo	r_4 = 0,10
viajou	r_2 = 0,09
a	r_5 = 0,09
noite	r_6 = 0,01
Ele	r_1 = 0,01
.	r_7 = 0,00

A implementação da Oclusão neste trabalho ordena todas as palavras de uma frase em relação ao impacto de sua remoção no valor da predição y da palavra alvo, que pode variar de 0 à 1. Esse valor é calculado como sendo o módulo da diferença entre os valores antes e depois da aplicação da técnica e à esse valor é dado o nome de Relevância R, também de 0 à 1, conforme intervalos de valores da própria predição Eq. 2.1.

$$R_{ki} = |y_{k0} - y_{ki}| \tag{2.1}$$

Sendo R_{ki} a Relevância da i_{esima} palavra em relação à palavra alvo k, y_{k0} o valor da predição da palavra alvo k e y_{ki} o valor da predição da palavra alvo k após a remoção da i_{esima} palavra.

Sendo NER uma aplicação da generalização de Hutchins (1987) automatizada por ML, a interpretação das Entidades conforme o contexto (instâncias) torna-se parte importante para uma classificação correta (SAMPSON, 1987). A Oclusão, dessa forma, pode permitir que as interações entre as palavras sejam visualizadas na forma de quais foram mais relevantes para a predição, servindo de explicação para usuários e de ponto de partida para possíveis melhorias na IA.

3 Trabalhos Relacionados

O aumento da demanda por XAI vem como consequência do impacto do uso de ML e IA (Inteligência Artificial) em variadas áreas, desde rotinas industriais e à tarefas pessoais (ADADI; BERRADA, 2018). Isso ocorre devido ao fato das pesquisas nessas áreas serem focadas na melhoria da predição, deixando cada vez mais complexa a compreensão do resultado transformando os modelos em verdadeiras caixas pretas de solução.

Em NLP, com o advento das ferramentas de busca em larga escala na *web*, estamos sendo confrontados com cada vez mais informação em texto (BRIN; PAGE, 1998), mas que, nem sempre, podemos compreender seu valor potencialmente escondido, conforme afirma Siddharthan (2002):

> *"Increasingly, businesses, government agencies and individuals are confronted with large amounts of text that are critical for working and living, but not well enough understood to get the enormous value out of them that they potentially hide."*

Esse fenômeno impulsionou pesquisas nas diversas áreas de NLP, como Análise de Sentimentos (YANG et al., 2019), Classificação Sintática (PoS Tagging do inglês, *Part-of-Speech Tagging*) (BOHNET et al., 2018) e NER (VAJJALA; BALASUBRAMANIAM, 2022) no sentido de melhorar a predição dos modelos, tal que NER vem sendo cada vez mais estudada devido ao seu uso em Recuperação de Informação *web*(LIU; CHEN; XIA, 2022).

Em seu trabalho, Vajjala e Balasubramaniam (2022) apresenta as principais arquiteturas estados da arte em NER. O que essas arquiteturas tem em comum, o que é recorrente em IA e ML, são suas complexas redes das quais as repostas são praticamente impostas ao usuário final humano. Adadi e Berrada (2018) discorre em seu trabalho sobre o avanço das técnicas de predição em detrimento da transparência das mesmas devido aos seus cálculos não lineares (SAMEK; WIEGAND; MÜLLER, 2017).

Por consequência do crescente uso dessas técnicas, diversas agencias governamentais tem aprovado leis de Proteção de Dados (GDPR, do inglês, *General Data Protection Regulation*) nas quais regras para a utilização de tomadas de decisões automáticas devem prover ao usuário direito de obter explicações significativas sobre as lógicas de decisão (GOODMAN; FLAXMAN, 2017; MALGIERI; COMANDÉ, 2017; WACHTER; MITTELSTADT; FLORIDI, 2017).

Um risco inerente ao uso de tecnologias de predição automática é a possibilidade de se haver tomadas de decisões equivocadas, podendo levar, inclusive, à fatalidades, como

acontecido em março de 2018 no Arizona (EUA)[1] (GUIDOTTI et al., 2018). Além disso, Pedreshi, Ruggieri e Turini (2008) observou, com o uso de Mineração de Dados, que decisões baseadas em Base de Dados, possuem tendencia de ser discriminatórias, podendo ser passadas ao modelos de ML durante o treinamento.

As XAIs, portanto, podem mitigar ou resolver tais problemas segundo Adadi e Berrada (2018), seguindo a metodologia de justificar, controlar, melhorar e descobrir conforme a explicação dada à predição, permitindo corrigir eventuais falhas do modelo ou vícios presentes no banco de dados.

As principais 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; FENG; BOYD-GRABER, 2018). Harbecke (2021), em seu trabalho, propõe o uso da Oclusão (ROBNIK-ŠIKONJA; KONONENKO, 2008; ZEILER; FERGUS, 2014; ZINTGRAF et al., 2017) com Modelagem de Linguagem com o uso de Axiomas.

Em seu trabalho, 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 $n \rightarrow 1$ (n entradas para 1 saída) e NERs são $n \rightarrow n$ (n entradas para n saídas).

Harbecke (2021) aprofunda seu trabalho aplicando Oclusão com Modelagem de Linguagem em Análise de Sentimentos usando a Oclusão como suporte para Modelagem de Linguagem em tarefas supervisionadas permitindo a substituição de palavras que não estão no vocabulário.

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 dessa técnica em NER neste presente 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.

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

4 Metodologia

Este capítulo apresenta como foi realizado o presente estudo com o objetivo de verificar a viabilidade do uso da explicabilidade por Oclusão aplicada em NER. A Figura 9 ilustra o fluxo das atividades. Primeiramente, treinou-se um modelo biLSTM para identificação das entidades nomeadas. Tal modelo foi criado levando em consideração uma amostra do dataset utilizado (treino). Logo após, 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, quais eram as palavras relevantes para classificá-la em um determinado rótulo. Com isso, foram realizadas análises considerando as anotações e as relevâncias obtidas por meio da explicação.

O restante desse capitulo apresenta de forma detalhada cada etapa de execução do presente trabalho: (*i*) os motivos para escolha do dataset de NER (Seção 4.1); (*ii*) como foi criada a Amostra de Dados Anotada de Relevância (Seção 4.4.1); (*iii*) o modelo de biLSTM utilizado para a rotulação (Seção 4.2); (*iv*) as características da técnica de Oclusão para a aplicação em NER (Seção 4.3); (*v*) a avaliação da precisão da XAI e análise quanto à variação do modelo de ML e dos valores das Relevâncias (Seção 4.4).

Figura 9 – Fluxo da Metodologia

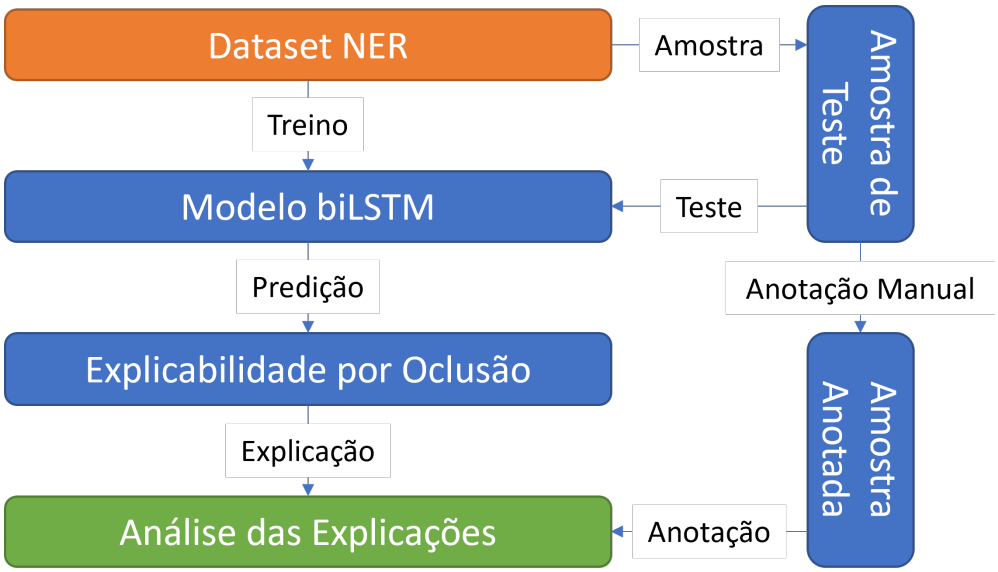


Tabela 5 – Distribuição do CoNLL 2003 inglês.

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

Fonte: Sang e Meulder (2003).

4.1 Base de Dados de NER

O dataset CoNLL 2003 (SANG; 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; STEVENSON; WHITEHEAD, 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 Paper with Code (2022)[1] apresenta dezenas de pesquisas que utilizam esse dataset tais como Wang et al. (2020), que usou arquiteturas LSTM e Transformers, citenschweter(Transformers) e Yamada et al. (2020) (Transformers), esses com valores de Precisão próximos de 94%.

O CoNLL 2003 também é a base de dados utilizada no treinamento do GloVe Embeddings (PENNINGTON; SOCHER; MANNING, 2014) e do modelo de biLSTM deste trabalho. Serão considerados apenas os dados em inglês dessa base, cuja distribuição da rotulação encontra-se detalhada na Tabela 5.

Devido à quantidade de referências em inglês para Explicabilidade e sendo o CoNLL2003 um dos principais datasets para o estudo de NER, optou-se por usá-lo em detrimento de outros datasets em português. Tal que ficam o dataset do HAREM (Avaliação de Reconhecedores de Entidades Mencionadas da língua portuguesa) (FREITAS et al., 2010) e o LeNER (dataset com entidades do meio jurídico) (ARAUJO et al., 2018) como sugestão para trabalhos futuros.

4.2 Modelo de Aprendizado de Máquina (biLSTM)

Neste trabalho, buscou-se trabalhar com dataset e modelo de Aprendizado de Máquina também utilizados em outras fontes de consulta. Como citado na Seção 2.2, este trabalho se baseou na implementação de Hu (2018). Seu trabalho realizou estudos com a XAI LRP aplicada para NER utilizando um modelo com a arquitetura biLSTM. Seu

[1] <https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003>

Tabela 6 – 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%

modelo utilizou o GloVe Embeddings (PENNINGTON; SOCHER; MANNING, 2014) e CoNLL 2003 Dataset (SANG; MEULDER, 2003) no treinamento. Esse modelo e dataset (Seção 4.1) serão usados como base do presente trabalho.

A arquitetura biLSTM apresentada por Graves e Schmidhuber (2005), como citado na Seção 2.2, tem a capacidade de rotular um dado conforme a sua vizinhança à esquerda ou à direita, assim como consegue trabalhar com tamanho de frases diversas. Por essas características, a biLSTM se tornou uma das principais arquiteturas utilizadas na predição em problemas de NER. Vajjala e Balasubramaniam (2022), em seu estudo sobre o Estado da Arte em NER de 2022, cita diversas variações dessa arquitetura com resultados acima de 90% de Precisão.

A partir do modelo pre-treinado, foram treinados cinco modelos diferentes, aumentando gradualmente o número épocas. Em uma Rede Neural, em cada época os pesos do modelo são ajustados de acordo com uma função de otimização[2,3], 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 nomedas, a Tabela 6 apresenta os valores da função de *Loss* e acurácia do treino e da validação. O conjunto de teste foi avaliado individualmente conforme amostras utilizadas na Seção 4.4.1. Além disso, é possível visualizar as Matrizes de Confusão de cada um dos modelos no Apêndice B. Como pode-se observar, os modelos apresentaram uma taxa de acurácia alta (acima dos 95%) tanto no treino quanto na validação.

Quando um modelo apresenta uma acurácia decrescente na validação, isso pode indicar que ele está em *overfitting* (superajuste), perdendo a capacidade de generalizar. Observa-se que os modelos apresentados na Tabela 6 não estão em *overfitting* pois não

[2] A função de otimização utilizada nos treinos foi a Entropia Cruzada.
[3] Função de Perda - Entropia Cruzada: Ela é calculada pela diferença entre duas funções de distribuição de probabilidade.

houve uma grandes variações nos valores das acurácias dos modelos.

Esses modelos foram utilizados para realizar as predições das Entidades nas diversas análises propostas cujo resultados estão apresentados no Capítulo 5.

4.3 Implementação da Oclusão

Conforme descrido na Seção 2.3, as técnicas de XAI podem se utilizar de estratégias de **Engenharia Reversa** ou **Perturbação de Instância** que pode ser realizada, dentre outras maneiras, por meio da Perturbação da Vizinhança. No presente trabalho, utilizou-se uma estratégia de Pertubação da vizinhança, no qual apenas alguns elementos (ou palavras) próximos do alvo são alterados na instância.

É importante ressaltar que esse trabalho tem por objetivo avaliar apenas a Explicação do Modelo e a Fidelidade do explicador ao modelo de ML utilizado. Para tal, foi implementada uma variação da técnica de Oclusão de Robnik-Šikonja e Kononenko (2008), usada em explicabilidade de imagens, para a aplicação em NER. Essa implementação remove cada palavra da sentença medindo a variação do valor da predição para rotulação da palavra alvo na saída, conforme descrição na Seção 2.4.

4.4 Análise dos Resultados

A Oclusão gera um ranking de palavras relevantes ordenada por valores entre 0 e 1. Dessa forma, será analisado no presente trabalho tais valores de Relevância R e também como essas explicações se aproximam da explicação apontada por humanos.

Para isso, realizamos uma anotação manual dessa relevância e utilizamos métricas de avaliação detalhadas a seguir.

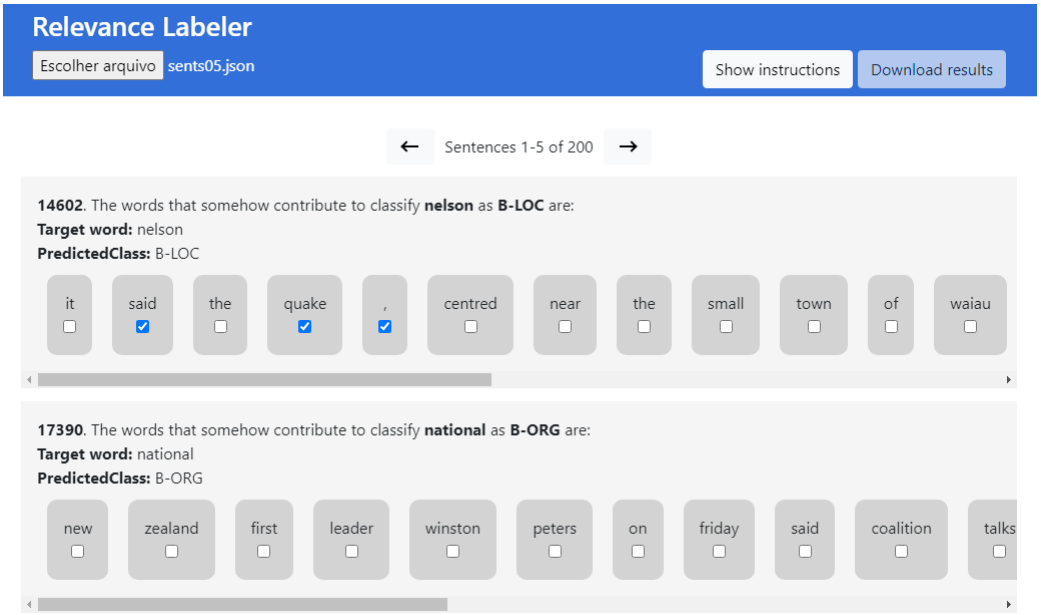
4.4.1 Amostra Anotada Manualmente com Palavras Relevantes

Foram selecionadas aleatoriamente 849 frases da base de dados adotada (CoNLL 2003) cujos requisitos para se selecionar uma frase são:

- Possuir mais que cinco palavras e menos que cinquenta;
- Possuir, ao menos, uma palavra rotulada com uma das categorias, **Pessoa**, **Localização** ou **Organização**.

A quantidade de palavras de cada rótulo pode ser observada na Tabela 7.

Figura 10 – Rotulador *Web* Desenvolvido no Escopo deste Trabalho.



Logo após, foi realizada a anotação desses dados com o auxílio de voluntários. Tais voluntários pertencem a grupos de estudantes de nível técnico, graduação e mestrado tanto de áreas de exatas, como tecnologia da informação e humanas.

Um comitê de três voluntários composto pelos próprios voluntários decidiu, por de maioria simples, 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 apresentado na Figura 10, na qual basta o usuário selecionar as caixa as quais ele considera relevante. Esta aplicação foi desenvolvida apenas com o intuito de selecionar um subconjunto de palavras do conjunto de palavras do texto. As explicações inseridas referem-se ao conjunto de dados de CoNLL, mas que pode ser facilmente alterado para outros datasets.

Nessa aplicação, o usuário faz o upload do arquivo com as sentenças que deseja rotular e, no navegador, clica nas palavras as quais considera relevante para a classificação da palavra alvo. Após a rotulação, o usuário faz o download do resultados que estarão representados por uma lista de verdadeiro ou falso conforme caixa de seleção marcada pelo usuário. Essa aplicação auxiliou os voluntários permitindo maior celeridade nessa tarefa.

Como conceito de relevância é subjetivo, antes da classificação foram apresentados

Tabela 7 – Distribuição da Amostra do CoNLL 2003 Anotada Manualmente.

Dataset	Frases	Palavras	PER	LOC	ORG	MISC
Amostra	849	849	126	420	303	0

exemplos do que pode ser considerado relevante no contexto do presente trabalho da seguinte forma:

- Palavras que possuem correlações gramaticais como concordâncias Verbais e Nominais:
  - Ex.: O verbo 'ir' indica que uma Pessoa vai para algum lugar (Local ou Organização);
  - Ex.: O adjetivo 'inteligente' é fortemente relacionado com Pessoa.
- Palavras com interdependências como Nomes e Sobrenomes:
  - Ex.: Se o Nome é de um Local, o Sobrenome também é de um Local;
  - Ex.: 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').
- Outras tal qual o voluntário entender, conforme contexto, que colaboraram no entendimento da rotulação.

4.4.2 Métrica de Avaliação

A presente avaliação permitirá verificar se a técnica é realmente capaz de reproduzir o entendimento de Relevância de cada palavra na rotulação de uma palavra alvo. Para isso, foi realizado (a) a análise da Precisão, considerando a anotação realizada pelo usuário (Seção 5.1); (b) um estudo de quais palavras a técnica de oclusão julgou como relevante (Seção 5.2); (c) uma análise do impacto da Relevância em modelos diferentes (Seção 5.3) e, finalmente, (d) uma análise de casos específicos de erros e acertos do método biLSTM e da Explicação(Seção 5.4).

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, conforme apresentado na Eq. 4.1.

$$\overline{P@k} = \frac{\sum_{i=0}^{N} P@k_i}{N} \tag{4.1}$$

Em que N é a quantidade de frases e $P@k_i$ representa a Precisão da Explicação para i_esima frase considerando as primeiras k posições. Tal Precisão é calculada como a proporção da quantidade de palavras relevantes até a k_esima posição do ranking.

Uma análise das explicações por Oclusão utilizando os modelos de ML treinados com épocas diferentes, apresentada na Seção 5.3 permitirá perceber a sensibilidade da

Relevância das palavras quanto ao impacto na substituição do modelo e, por fim, serão apresentado na Seção 5.4 estudos de casos onde existem falhas na explicação em relação às anotações e casos onde existem falhas de predição. O estudo de caso das explicações quanto ao erro de predições visa verificar o que ocorre quando existe um erro de predição na rotulação e discutir possíveis estratégias para mitigar esse erro. Além disso, tais casos podem sinalizar direções para pesquisas de melhorias no ML utilizando as XAIs.

5 Resultados

Para avaliar a explicação de uma XAI é preciso compreender a forma que a técnica cria sua explicação. A Oclusão simula alterações na entrada do modelo (Perturbação na Vizinhança) medindo o impacto dessas mudanças no valor da predição. Por outro lado, o modelo de ML é quem possui os pesos e gera os resultados das predições que serão utilizados para explicação. Esse entendimento é o ponto de partida para se avaliar uma explicação tal que, algumas predições poderão apresentar explicações imprecisas ou que não fazem sentido para o ser humano, mas que trazem consigo informações nos pesos do modelo de ML que evidenciam interações entre palavras que passariam despercebidas por humanos (GUIDOTTI et al., 2018; ADADI; BERRADA, 2018).

Por esse motivo, os modelos não necessariamente irão considerar relevantes conceitos gramaticais como concordâncias. Portanto, diferentes aspectos influenciam no resultado do aprendizado como: *under* e *over fitting*, o balanceamento do banco de dados e até o *Embedding* podem apresentar influência no resultado (GUIDOTTI et al., 2018; ADADI; BERRADA, 2018).

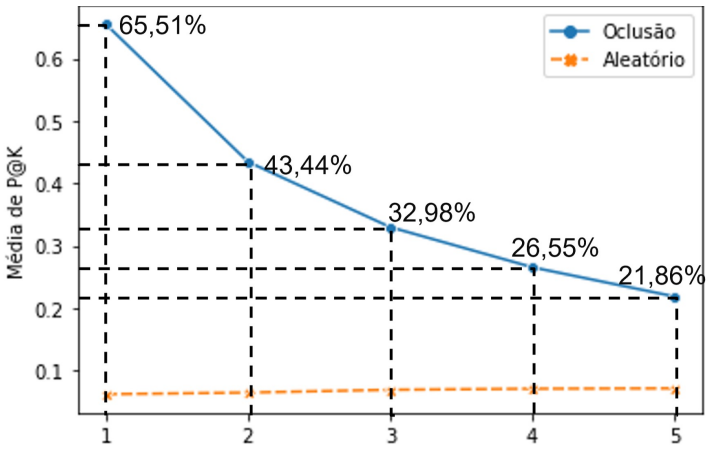
Dessa forma, na Seção 5.1 as explicações foram comparadas quanto ao entendimento humano utilizando-se como *Ground Truth* as amostras anotadas manualmente descritas na Seção 4.4.1; na Seção 5.3 foram feitas análises mais aprofundada da relevancia predita e anotada pelo usuário por meio da distribuição dos valores de relevâncias gerados pela técnica de explicabilidade e comparados com a anotação de manual; na Seção 5.3 foram analisados os impactos da mudança dos modelos de ML nas explicações da Oclusão; e na Seção 5.4 foram realizados estudos de casos em busca de possibilidades de pesquisas futuras para melhoria na predição do modelo de ML e nas XAIs.

5.1 Análise 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 11 apresenta, pra cada posição k a sua precisão média ($\overline{P@k}$), cf. Eq. 4.1. 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.

Pode-se verificar na Figura 11 que a $\overline{P@k}$ diminui à medida que se aumenta o valor de k, se aproximando de uma seleção aleatória de valores de relevância. Esse

Figura 11 – Precisão



comportamento reflete o aumento da quantidade das palavras testadas e, com isso, o aumento na probabilidade de erro. 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 5.2.

Nesse Gráfico, Figura 11, os valores da Precisão para k entre 1 e 3 permanecem 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 da frase. De forma análoga para k entre 1 e 4, acima de 25% e entre 1 e 5, acima de 20%.

5.2 Análise dos Valores de Relevância Estimados pela Oclusão

Para cada frase, a técnica de Oclusão retorna um valor contínuo atribuído a cada palavra da frase em relação à palavra a ser classificada. Com isso, obtemos um ranqueamento do quão Relevante as palavras de uma frase são na classificação de uma entidade alvo.

No entanto, o conceito de ser ou não Relevante adotado na Seção 4.4.1 na escolha das palavras é discreto, o que significa que uma palavra é ou não Relevante, independente da sua posição no rank. Dessa forma, esta Seção busca analisar o valor da Relevância e compará-los à relevância apontada por um humano na anotação manual. Além disso, essa 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 (cf. Seção 2.4).

A Figura 12 levou em consideração a distribuição da relevância das 5 palavras mais bem ranqueadas pela Oclusão em cada frase. Sendo que na Figura 12(a) considerou-se apenas as palavras Relevantes (anotadas pelos voluntários) e na Figura 12(b), as não

relevantes (não anotadas pelos voluntários). Os valores de relevância destes gráficos estão expressos por ordem de grandeza na escala logarítmica (10^{-1}, 10^{-2}, 10^{-3}...).

Percebe-se, na Figura 12(a) que a ordem de grandeza acima de 10^{-1} possui aproximadamente $37,54\%$ palavras relevantes anotadas, e acima de 10^{-2} aproximadamente $67,51\%$. 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 12(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 12(a), podendo-se considerar as demais como desnecessárias na dialética humana.

Para compreender melhor a diferença entre o rank ordenado por Oclusão e o entendimento humano da Relevância, os valores utilizados na Figura 12(a, b) foram separados por posição no rank, ainda obedecendo a escala logarítmica, e apresentados na Figura 13(a, b) tal que as curvas indicam as distribuições por densidade de palavras por ordem de grandeza em cada rank dos Top 5 melhores ranqueados. Similarmente aos histogramas, na Figura 13(a) estão apenas as palavras que foram anotadas como relevantes e na Figura 13(b) seu complemento.

Verifica-se, na Figura 13(a), que a curva do rank 1 apresenta-se bem mais acentuada que as demais com 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 13(b). Os demais ranks possuem curvas bem menos inclinadas na Figura 13(a), diminuindo conforme aumenta a posição no rank, mostrando que quanto menor o rank, menor a chance de um humano considerar a palavra relevante, além disso, os picos dessas curvas vão se deslocando no sentido de ordem de grandeza menores, permitindo concluir que o valor de R reflete o comportamento da precisão por posição do rank. A Figura 13(b) evidencia mais essa tendência de erro a medida em que o valor de R e a

Figura 12 – Distribuição das Top5 Palavras Relevantes Anotadas por Ordem de Grandeza.

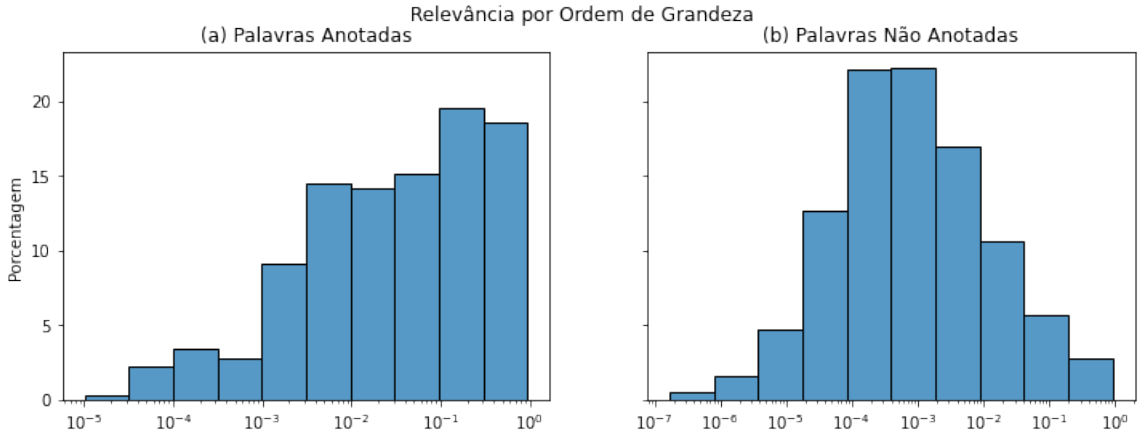
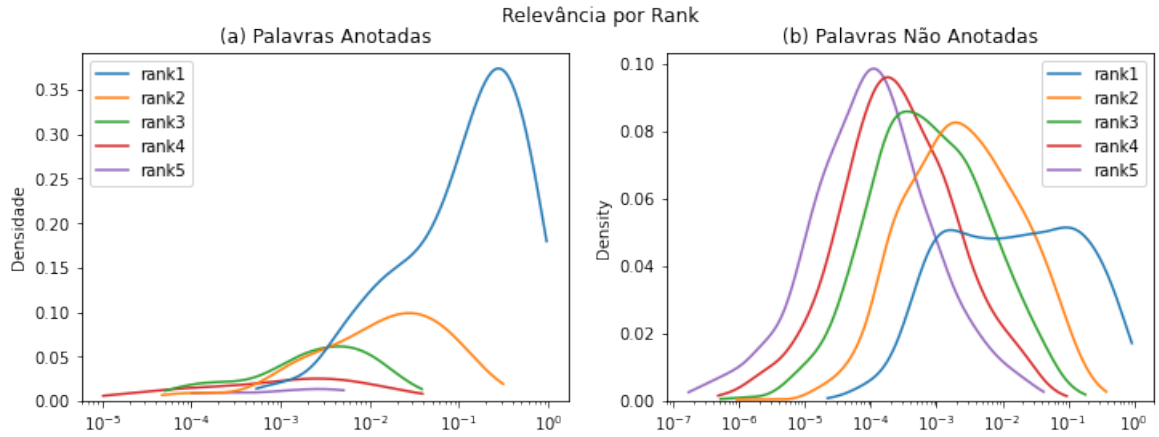


Figura 13 – Densidade das Top5 Palavras Relevantes Anotadas e Não Anotadas por Rank.



posição da palavra no rank diminuem.

Como a Oclusão não considera a própria palavra no ranqueamento, devido a forma de calculo 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 pois as demais contribuíram muito pouco para tal. Por exemplo, Fifa (Federação Internacional de Futebol) é sempre 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.

5.3 Análise do Impacto do Modelo de Aprendizado de Máquina na Explicabilidade

Nesta seção será verificado como alterações do modelo interferem na Explicação. Para tal, foram criados modelos variando-se o treinamento. Vale ressaltar que esses modelos foram pré-treinados tal que o treinamento apresentado na Tabela 8 pode ser considerado além do necessário. As amostras de teste foram avaliadas individualmente conforme seleção apresentada na Seção 4.4.1. Tais amostras foram preditas pelos cinco modelos e seus resultados submetidos à explicação por Oclusão. Esse resultado permitirá analisar a variação do valor de Relevância para as palavras nessas frases, considerando os modelos

Tabela 8 – 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%

treinados com épocas diferentes.

As amostras descritas na Seção 4.4.1 preditas pelos cinco modelos e seus resultados submetidos à explicação por Oclusão. Esse resultado nos permitirá analisar a variação do valor de Relevância para as palavras nessas frases, considerando os modelos treinados com épocas diferentes. Para uma análise mais detalhada, a Figura 14 apresenta a distribuição da palavra mais bem ranqueada em cada frase em cada modelo. Logo após, a Figura 15(a, b, c, d) mostra também as palavras da 2ª a 5ª posição no ranking (rank 2 a 5). Os valores de Relevância R são apresentados em escala logarítmica.

Na Figura 14, 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 cada vez menores até que no Modelo 5 elas apresentam-se em maior concentração em 10^{-4}.

Nesses casos, onde os valores de Relevância são muito pequenos, pode-se compreender que o modelo esta superajustado para predição da palavra alvo sendo a mesma

Figura 14 – Distribuição dos Valores por Modelo - Rank 1.

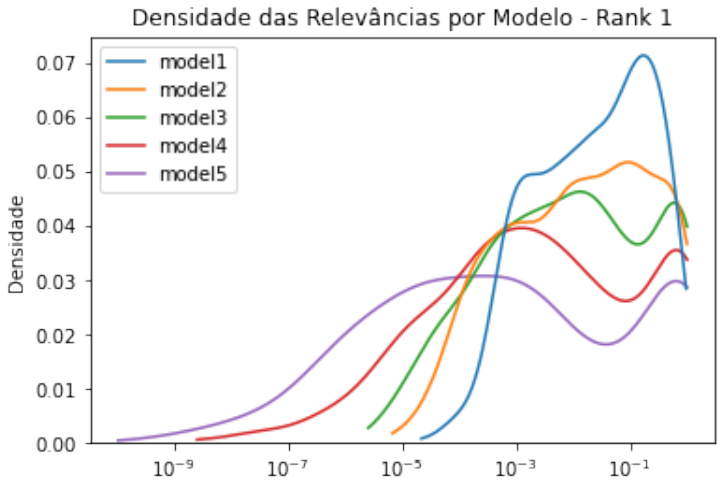
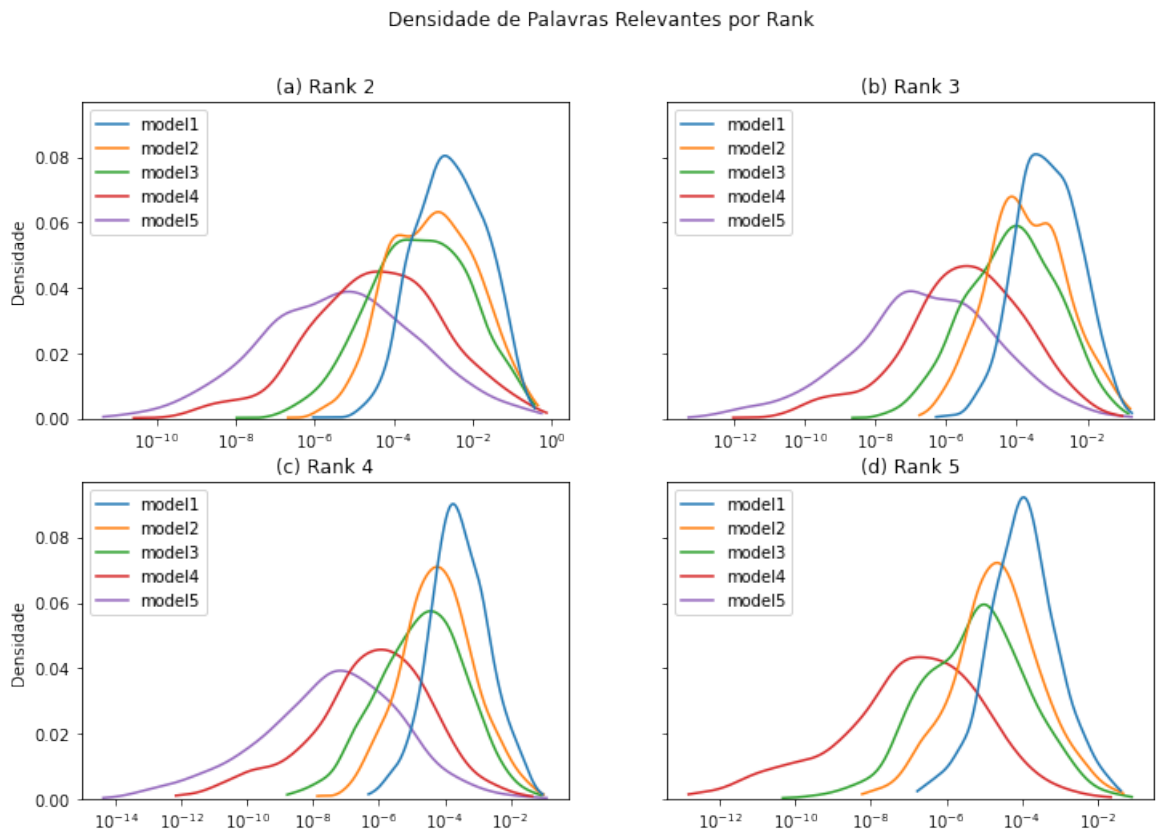


Figura 15 – Distribuição dos Valores por Modelo - Ranks 2, 3, 4, 5.



autosuficiente para sua classificação. Esta seria uma indicação indireta de Relevância da palavra alvo sobre ela mesma, conforme visto na Seção 5.2 e evidenciada pela análise destes modelos.

O fenômeno da Figura 14 se replica nas Figuras 15(a), 15(b), 15(c), 15(d) com o valor médio de R se deslocando de 10^{-2} para ordens menores que 10^{-5}. Quando a Oclusão não considera palavra vizinha alguma como relevante, pode-se concluir que o modelo considerou apenas a própria palavra alvo para a sua predição. Dessa forma, podemos perceber que há um ajuste cada vez maior em relação à palavra alvo a medida que aumenta-se o número de épocas.

5.4 Estudos de Casos das Explicações por Oclusão

Nessa seção serão apresentadas algumas explicações buscando verificar se (i) elas estão cumprindo seu objetivo; (ii) o que ocorre quando a explicação não reflete a compreessão humana; e (iii) o que a técnica explica em caso de falha na predição. Para equalizar as análises, apenas o Modelo 1 foi utilizado.

Tabela 9 – Predições por Modelo -Estudo de Caso 1.

	Word	True	-	Pred1	-	Pred2	-	Pred3
0-	there	O	-	O	-	O	-	O
1-	were	O	-	O	-	O	-	O
2-	apparently	O	-	O	-	O	-	O
3-	no	O	-	O	-	O	-	O
4-	adults	O	-	O	-	O	-	O
5-	at	O	-	O	-	O	-	O
6-	the	O	-	O	-	O	-	O
7-	party	O	-	O	-	O	-	O
8-	as	O	-	O	-	O	-	O
9-	the	O	-	O	-	O	-	O
10-	father	O	-	O	-	O	-	O
11-	of	O	-	O	-	O	-	O
12-	the	O	-	O	-	O	-	O
13-	family	O	-	O	-	O	-	O
14-	who	O	-	O	-	O	-	O
15-	lived	O	-	O	-	O	-	O
16-	in	O	-	O	-	O	-	O
17-	the	O	-	O	-	O	-	O
18-	house	O	-	O	-	O	-	O
19-	was	O	-	O	-	O	-	O
20-	out	O	-	O	-	O	-	O
21-	of	O	-	O	-	O	-	O
22-	town	O	-	O	-	O	-	O
23-	and	O	-	O	-	O	-	O
24-	the	O	-	O	-	O	-	O
25-	mother	O	-	O	-	O	-	O
26-	died	O	-	O	-	O	-	O
27-	more	O	-	O	-	O	-	O
28-	than	O	-	O	-	O	-	O
29-	a	O	-	O	-	O	-	O
30-	year	O	-	O	-	O	-	O
31-	ago	O	-	O	-	O	-	O
32-	,	O	-	O	-	O	-	O
33-	black	B-PER	-	O	-	O	-	O
34-	said	O	-	O	-	O	-	O
35-	.	O	-	O	-	O	-	O

5.4.1 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*, cuja rótulo seria **B-PER**, foi predita como **outro** pelos 3 modelos estudados (Modelo 1, Modelo 2 e Modelo 3 da Tabela 8), como pode-se verificar na Tabela 9.

Tabela 10 – Relevância Ordenadas 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)
5-	,	(2.58E-05)	said	(4.53E-05)	.	(4.36E-05)
6-	.	(7.98E-06)	house	(3.88E-05)	died	(2.95E-05)
7-	died	(2.16E-06)	family	(1.24E-05)	father	(2.52E-05)
8-	the	(1.73E-06)	father	(8.67E-06)	said	(2.02E-05)
9-	lived	(1.40E-06)	who	(4.53E-06)	who	(1.08E-05)
10-	family	(1.12E-06)	lived	(1.37E-06)	family	(6.02E-06)
11-	who	(9.49E-07)	party	(1.19E-06)	of	(5.90E-06)
12-	year	(9.00E-07)	the	(1.09E-06)	lived	(3.54E-06)
13-	father	(6.67E-07)	died	(1.08E-06)	than	(2.48E-06)
14-	of	(3.18E-07)	year	(1.08E-06)	party	(2.36E-06)
15-	the	(1.40E-07)	the	(8.20E-07)	more	(1.74E-06)
16-	town	(1.34E-07)	of	(4.96E-07)	the	(1.36E-06)
17-	and	(1.28E-07)	than	(2.37E-07)	as	(1.21E-06)
18-	a	(1.07E-07)	the	(2.34E-07)	year	(8.71E-07)
19-	the	(1.02E-07)	a	(1.30E-07)	there	(7.33E-07)
20-	party	(6.30E-08)	more	(1.29E-07)	no	(4.95E-07)
21-	out	(6.18E-08)	the	(9.29E-08)	the	(2.83E-07)
22-	more	(4.07E-08)	of	(8.73E-08)	adults	(2.82E-07)
23-	of	(2.70E-08)	there	(7.18E-08)	of	(2.28E-07)
24-	than	(2.36E-08)	and	(6.95E-08)	out	(2.27E-07)
25-	apparently	(9.36E-09)	as	(6.94E-08)	a	(1.90E-07)
26-	was	(8.02E-09)	apparently	(5.90E-08)	apparently	(1.84E-07)
27-	there	(6.66E-09)	the	(5.81E-08)	the	(1.39E-07)
28-	the	(6.62E-09)	adults	(5.15E-08)	the	(5.75E-08)
29-	the	(3.27E-09)	at	(1.03E-08)	at	(4.97E-08)
30-	in	(3.03E-09)	town	(8.84E-09)	the	(3.28E-08)
31-	no	(2.61E-09)	no	(5.99E-09)	town	(2.74E-08)
32-	adults	(2.09E-09)	out	(5.41E-09)	was	(1.32E-08)
33-	as	(1.55E-09)	was	(3.13E-09)	and	(1.20E-08)
34-	at	(4.40E-10)	were	(3.03E-09)	were	(4.64E-09)
35-	were	(2.39E-10)	in	(2.18E-09)	in	(3.41E-09)

A Tabela 10 apresenta a ordenação das Relevância deste caso. Nas explicações dadas aos 3 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 ficará comprometida.

Tabela 11 – Predições por Modelo -Estudo de Caso 2.

	Word	True	-	Pred1	-	Pred2	-	Pred3
0-	1.	O	-	O	-	O	-	O
1-	tom	B-PER	-	B-PER	-	B-PER	-	B-PER
2-	pukstys	I-PER	-	I-PER	-	I-PER	-	I-PER
3-	(	O	-	O	-	O	-	O
4-	u.s.	B-LOC	-	B-LOC	-	B-LOC	-	B-LOC
5-	)	O	-	O	-	O	-	O
6-	86.82	O	-	O	-	O	-	O

Tabela 12 – Relevância Ordenadas 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)
5-	u.s.	(2.76E-05)	u.s.	(1.33E-05)	(	(1.36E-06)
6-	)	(2.99E-06)	)	(4.58E-07)	)	(6.26E-08)

5.4.2 Erro de Explicação

Na predição da sentença *"1. tom pukstys (u.s.) 86.82"* a palavra alvo *tom*, cuja rótulo é **B-PER**, foi predita corretamente com valor de $y > 9 * 10^{-1}$ pelos 3 modelos estudados, como pode-se verificar na Tabela 11, contudo, a palavra mais relevante no rank não foi anotado (palavra "1"), a invés disso, a palavra *pukstys* estava anotada.

Na Tabela 12 pode-se perceber que o valor da Relevância R para ambas as palavras ficou na ordem de 10^{-1}, o que indica que ambas foram relevantes na predição da palavra alvo. Contudo, apenas a palavra *pukstys* estava anotada, esta, provavelmente, por ser o sobrenome do *tom* (*pukstys* é um **I-PER**), o que induz o humano à considerá-la como relevante.

5.4.3 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 3 modelos estudados, como pode-se verificar na Tabela 13, contudo, não existem palavras anotadas tal que os voluntários considerassem que alguma outra palavra seria relevante para a classificação de *mills*.

Na Tabela 14 pode-se verificar que o valor da Relevância R abaixo de 10^{-2}, o que

Tabela 13 – Predições por Modelo -Estudo de Caso 3.

	Word	True	-	Pred1	-	Pred2	-	Pred3
0-	as	O	-	O	-	O	-	O
1-	glenn	B-PER	-	B-PER	-	B-PER	-	B-PER
2-	lawhon	I-PER	-	I-PER	-	I-PER	-	I-PER
3-	,	O	-	O	-	O	-	O
4-	a	O	-	O	-	O	-	O
5-	rural	O	-	O	-	O	-	O
6-	florida	B-LOC	-	B-LOC	-	B-LOC	-	B-LOC
7-	minister	O	-	O	-	O	-	O
8-	who	O	-	O	-	O	-	O
9-	is	O	-	O	-	O	-	O
10-	the	O	-	O	-	O	-	O
11-	victim	O	-	O	-	O	-	O
12-	's	O	-	O	-	O	-	O
13-	father	O	-	O	-	O	-	O
14-	,	O	-	O	-	O	-	O
15-	looked	O	-	O	-	O	-	O
16-	on	O	-	O	-	O	-	O
17-	,	O	-	O	-	O	-	O
18-	mills	B-PER	-	B-PER	-	B-PER	-	B-PER
19-	was	O	-	O	-	O	-	O
20-	pronounced	O	-	O	-	O	-	O
21-	dead	O	-	O	-	O	-	O
22-	at	O	-	O	-	O	-	O
23-	7:13	O	-	O	-	O	-	O
24-	a.m.	O	-	O	-	O	-	O
25-	est	O	-	O	-	O	-	O
26-	(	O	-	O	-	O	-	O
27-	1213	O	-	O	-	O	-	O
28-	gmt	B-MISC	-	B-MISC	-	B-MISC	-	B-MISC
29-	)	O	-	O	-	O	-	O
30-	for	O	-	O	-	O	-	O
31-	the	O	-	O	-	O	-	O
32-	murder	O	-	O	-	O	-	O
33-	of	O	-	O	-	O	-	O
34-	lester	B-PER	-	B-PER	-	B-PER	-	B-PER
35-	lawhon	I-PER	-	I-PER	-	I-PER	-	I-PER
36-	.	O	-	O	-	O	-	O

indica que **mills** foi suficiente para se rotular como **B-PER**. Este é um dos casos de Explicação indireta onde a explicação depende da insuficiência de valores de Relevância das demais palavras.

Tabela 14 – Relevância Ordenadas 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)
5-	lawhon	(2.17E-04)	lester	(1.34E-04)	florida	(4.01E-05)
6-	lester	(2.16E-04)	7:13	(6.46E-05)	who	(3.21E-05)
7-	,	(8.67E-05)	lawhon	(5.18E-05)	looked	(2.67E-05)
8-	est	(4.78E-05)	1213	(4.94E-05)	's	(2.33E-05)
9-	on	(3.03E-05)	dead	(3.76E-05)	rural	(1.09E-05)
10-	was	(2.34E-05)	florida	(2.98E-05)	was	(8.97E-06)
11-	glenn	(2.17E-05)	was	(2.72E-05)	victim	(6.55E-06)
12-	minister	(2.03E-05)	lawhon	(2.09E-05)	glenn	(4.17E-06)
13-	a.m.	(1.89E-05)	minister	(1.77E-05)	,	(3.05E-06)
14-	florida	(1.64E-05)	glenn	(1.53E-05)	lawhon	(2.33E-06)
15-	dead	(1.36E-05)	a.m.	(1.05E-05)	as	(2.30E-06)
16-	pronounced	(1.11E-05)	who	(1.05E-05)	of	(2.26E-06)
17-	at	(9.66E-06)	's	(8.00E-06)	is	(1.76E-06)
18-	who	(9.45E-06)	,	(6.19E-06)	1213	(1.54E-06)
19-	's	(9.31E-06)	victim	(5.71E-06)	lawhon	(1.51E-06)
20-	lawhon	(8.89E-06)	est	(4.57E-06)	dead	(1.42E-06)
21-	murder	(2.74E-06)	at	(3.37E-06)	minister	(1.13E-06)
22-	victim	(2.48E-06)	rural	(1.47E-06)	7:13	(9.98E-07)
23-	as	(1.56E-06)	on	(1.09E-06)	a.m.	(9.35E-07)
24-	rural	(9.11E-07)	is	(5.17E-07)	a	(4.10E-07)
25-	the	(7.82E-07)	the	(3.25E-07)	at	(2.31E-07)
26-	looked	(5.65E-07)	gmt	(2.91E-07)	est	(1.71E-07)
27-	gmt	(4.98E-07)	(	(2.30E-07)	murder	(1.66E-07)
28-	is	(3.57E-07)	murder	(2.30E-07)	the	(1.36E-07)
29-	(	(3.05E-07)	as	(1.44E-07)	gmt	(9.88E-08)
30-	.	(1.16E-07)	for	(9.95E-08)	for	(8.26E-08)
31-	for	(6.43E-08)	a	(8.09E-08)	on	(7.86E-08)
32-	a	(6.42E-08)	of	(6.20E-08)	the	(7.01E-08)
33-	of	(5.63E-08)	)	(1.23E-08)	(	(1.38E-08)
34-	,	(4.26E-08)	the	(8.31E-09)	)	(2.47E-09)
35-	)	(4.25E-08)	,	(6.25E-09)	,	(1.60E-09)
36-	the	(6.75E-09)	.	(3.99E-09)	.	(1.36E-09)

6 Conclusão

Neste presente trabalho, buscou-se analisar a capacidade da Explicabilidade por Oclusão de explicar resultados na classificação de palavras por rotulador automatizado em tarefas de NER. Os principais rotuladores automáticos utilizam-se de modelos de Aprendizado de Máquina, sendo esses de difícil entendimento do usuário final.

Dada a crescente demanda de se explicar resultados de IA e a complexidade dessa tarefa, é importante considerar abordagens que entreguem ao usuário final respostas mais contextualizadas e simples. Em especial, as tarefas em NLP, onde as explicações humanas estão relacionadas com entendimentos linguísticos, explicações que refazem as complexas camadas de Redes Neurais ou apresentadas de forma numérica podem ser pouco satisfatórias. Dessa forma a Oclusão em NER apresenta um algoritmo de fácil entendimento no qual suas explicações apontam relações simples entre palavras da mesma frase dando entendimento contextualizado.

A Oclusão obtém suas explicações por meio de Perturbação da Vizinhança de forma que ela remove palavras da frase com o objetivo de calcular qual a diferença do valor da predição da palavra alvo após essa remoção. Para verificar se a Oclusão estava oferecendo explicações compatíveis com o entendimento humanos, foi criado um banco de dados anotado com palavras que humanos consideravam 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 por humanos. 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. Nessa análise, buscou-se entender qual seria o valor da Relevância R, calculado pela Oclusão, tal que o humano entendia como Relevante. Nos resultados obtidos neste trabalhos, ficou evidenciado que existe um valor de corte tal que acima dele, a maior parte das palavras que a Oclusão considerou como Relevante, o humano considerou como Relevante. Essa análise permitiu que se entendesse, por exemplo, que se não existirem palavras no rank com valor de R acima do valor de corte, a própria palavra alvo pode estar sendo responsável por sua classificação.

Em um terceiro momento, foi analisado o que ocorre quando uma mesma instância é

submetida à modelos com pesos diferentes. Para tal, foi usado o mesmo modelo com apenas quantidade de épocas de treino diferentes. Nenhum dos modelos apresentou *overfitting*, contudo, ficou evidente, que quanto mais se treinou o modelo, mais bem ajustadas ficaram as predições da palavra alvo no sentido da mesma ficar cada vez mais autossuficiente para sua predição.

Estes resultados indicam que a Oclusão é uma técnica que permite que o usuário obtenha uma explicação para a classificada de uma palavra. Essa explicação foi muito superior a gerada de forma aleatória, seja pela indicação de pelo menos uma palavra que ajudou na predição ou seja pela falta de palavras acima do valor de corte, que pode indicar autossuficiência da palavra alvo, com a Precisão acima de 65% na indicação de, pelo menos, uma palavra relevante. Além disso, pode-se estudar casos de superajustes do modelos para determinadas instâncias, permitindo ao usuário visualizar estes casos com o auxílio da explicação. Compreensão do ranqueamento por Relevância ainda podem ajudar a verificar interações de palavras que não faz sentido de tal forma que o usuário pode recorrer a uma exploração dos pesos e/ou do banco de dados para entender o motivo dessa relação inesperada.

A Oclusão requer algumas análises secundárias que precisam ser discutidas e alvo de trabalhos futuros. Neste trabalho considerou-se apenas palavras individualizadas nos estudos apresentados. No entanto, sabe-se que na linguística existem palavras que possuem um significado completamente diferente de quando elas estão acompanhadas por outras palavras específicas, essas chamados de expressões. Para esses casos, a Oclusão pode ser refinada para abranger mais de uma palavra no calculo de Relevância, o que não foi abordado no presente trabalho.

Outra abordagem futura é o caso da palavra alvo depender apenas de si mesma para ser predita. Nesse trabalho, tal conclusão é obtida de maneira indireta, ou seja, pelo calculo da Relevância das demais palavras. A inferência que levou a essa suposição pode ser equivocada sendo esse um caso que deve ser melhor explorado no futuro.

Existe também uma necessidade de se aprofundar pesquisas para entender a variação do rank. Apesar de representar quais as palavras foram mais relevantes para a predição, em ordem do maior para o menor. Nas posições mais baixas a variação do valor de R são muito pequenas, podendo omitir erros de aproximação. Além disso, o ao valor de corte da Relevância traz consigo uma subjetividade não necessariamente ligada à posição no rank, o que pode induzir o usuário à acreditar que apenas as primeiras posições são relevantes.

Essas análises, juntamente com o banco de dados anotado manualmente e a aplicação *Web* ficam a disposição da comunidade como contribuição do presente trabalho podendo servir, inclusive, como referência de trabalhos futuros.

Trabalhos futuros no uso da Oclusão como explicador para NER, conforme mencio-

nado, devem considerar a possibilidade de estudar a ocorrência de expressões linguísticas e uma forma de calcular a Relevância delas. Harbecke (2021) usou em seu trabalho Oclusão com Modelagem de Linguagem para substituir palavras que não estavam presentes no vocabulário em entidade similares, essa abordagem talvez possa ser aplicada com êxito no contexto de NER para substituir expressões por palavras com similaridade de significado.

Quanto à palavras autossuficientes para sua predição, uma abordagem mais profunda pode ser importante para, ao invés de uma conclusão indireta, produzir valores de Relevância R calculados diretamente pela Oclusão.

E, finalmente, ampliar este estudo para outros bancos de dados ou até outras tarefas de NLP, como PoS *Tagging* pode permitir que mais usuários se beneficiem dos resultados obtidos no contexto do presente trabalho.

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, v. 6, p. 52138–52160, 2018.

ARAUJO, P. H. Luz de et al. Lener-br: a dataset for named entity recognition in brazilian legal text. In: SPRINGER. *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*. [S.l.], 2018. p. 313–323.

ARRAS, L. et al. Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*, 2017.

BOHNET, B. et al. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. *arXiv preprint arXiv:1805.08237*, 2018.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1, p. 107–117, 1998.

DANILEVSKY, M. et al. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

FREITAS, C. et al. Second harem: advancing the state of the art of named entity recognition in portuguese. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association*. [S.l.], 2010.

GOODMAN, B.; FLAXMAN, S. European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, v. 38, n. 3, p. 50–57, 2017.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, Elsevier, v. 18, n. 5-6, p. 602–610, 2005.

GRISHMAN, R.; SUNDHEIM, B. *Design of the MUC-6 evaluation*. [S.l.], 1996.

GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.

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

HEWITT, J.; MANNING, C. D. A structural probe for finding syntax in word representations. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4129–4138.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HU, J. *Explainable Deep Learning for Natural Language Processing*. 2018.

HUTCHINS, J. Summarization: Some problems and methods. *Meaning: The frontier of informatics*, Citeseer, v. 9, p. 151–173, 1987.

JONES, K. S. Automatic summarizing: factors immarizing: factors and directions. *Advances in automatic text summarization*, MIT press, p. 1, 1999.

KRIPKE, S.; DAVIDSON, D.; HARMAN, G. *Semantics of natural language*: Naming and necessity. 2. ed. Springer, Dordrecht: Springer Nature Switzerland AG, 1972. 253-355 p.

LENT, M. V.; FISHER, W.; MANCUSO, M. An explainable artificial intelligence system for small-unit tactical behavior. In: MENLO PARK, CA; CAMBRIDGE, MA; LONDON; AAAI PRESS; MIT PRESS; 1999. *Proceedings of the national conference on artificial intelligence*. [S.l.], 2004. p. 900–907.

LIU, J. U. K. J. W. *Deep learning for NLP and speech recognition.*: Text and speech basics. Cham, Switzerland: Springer Nature Switzerland AG, 2019. 87-134 p.

LIU, X.; CHEN, H.; XIA, W. Overview of named entity recognition. *Journal of Contemporary Educational Research*, v. 6, n. 5, p. 65–68, 2022.

MALGIERI, G.; COMANDÉ, G. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, Oxford University Press, v. 7, n. 4, p. 243–265, 2017.

PAICE, C. D.; JONES, P. A. The identification of important concepts in highly structured technical papers. In: *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval*. [S.l.: s.n.], 1993. p. 69–78.

PAPER WITH CODE. *Named Entity Recognition on CoNLL 2003 (English)*. [S.l.], 2022.

PEDRESHI, D.; RUGGIERI, S.; TURINI, F. Discrimination-aware data mining. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2008. p. 560–568.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543.

QI, P. T. D. Y. Z. C. D. M. Proceedings of the conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Computer networks and ISDN systems*, Association for Computational Linguistics, p. 160–170, 2018.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 20, n. 5, p. 589–600, 2008.

RUSSELL-ROSE, T.; STEVENSON, M.; WHITEHEAD, M. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. Goldsmiths, University of London, 2002.

SAMEK, W.; WIEGAND, T.; MÜLLER, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

SAMPSON, G. Probabilistic models of analysis. *The computational analysis of English*, Longman New York, v. 30, p. 41, 1987.

SANG, E. F.; MEULDER, F. D. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

SIDDHARTHAN, A. 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*, Cambridge University Press, v. 8, n. 1, p. 91–92, 2002.

TORFI, A. et al. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*, 2020.

VAJJALA, S.; BALASUBRAMANIAM, R. What do we really know about state of the art ner? *arXiv preprint arXiv:2205.00034*, 2022.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, Oxford University Press, v. 7, n. 2, p. 76–99, 2017.

WALLACE, E.; FENG, S.; BOYD-GRABER, J. Interpreting neural networks with nearest neighbors. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 136–144. Disponível em: <https://aclanthology.org/W18-5416>.

WANG, X. et al. Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*, 2020.

YAMADA, I. et al. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.

YANG, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, v. 32, 2019.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. [S.l.], 2014. p. 818–833.

ZINTGRAF, L. M. et al. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

Apêndices

APÊNDICE A – Loss e Acurácia dos Modelos Utilizados

Tabela 15 – 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%

APÊNDICE B – Matrizes de Confusão dos Modelos Utilizados

Figura 16 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 5 Épocas.

Figura 17 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 10 Épocas.

Figura 18 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 20 Épocas.

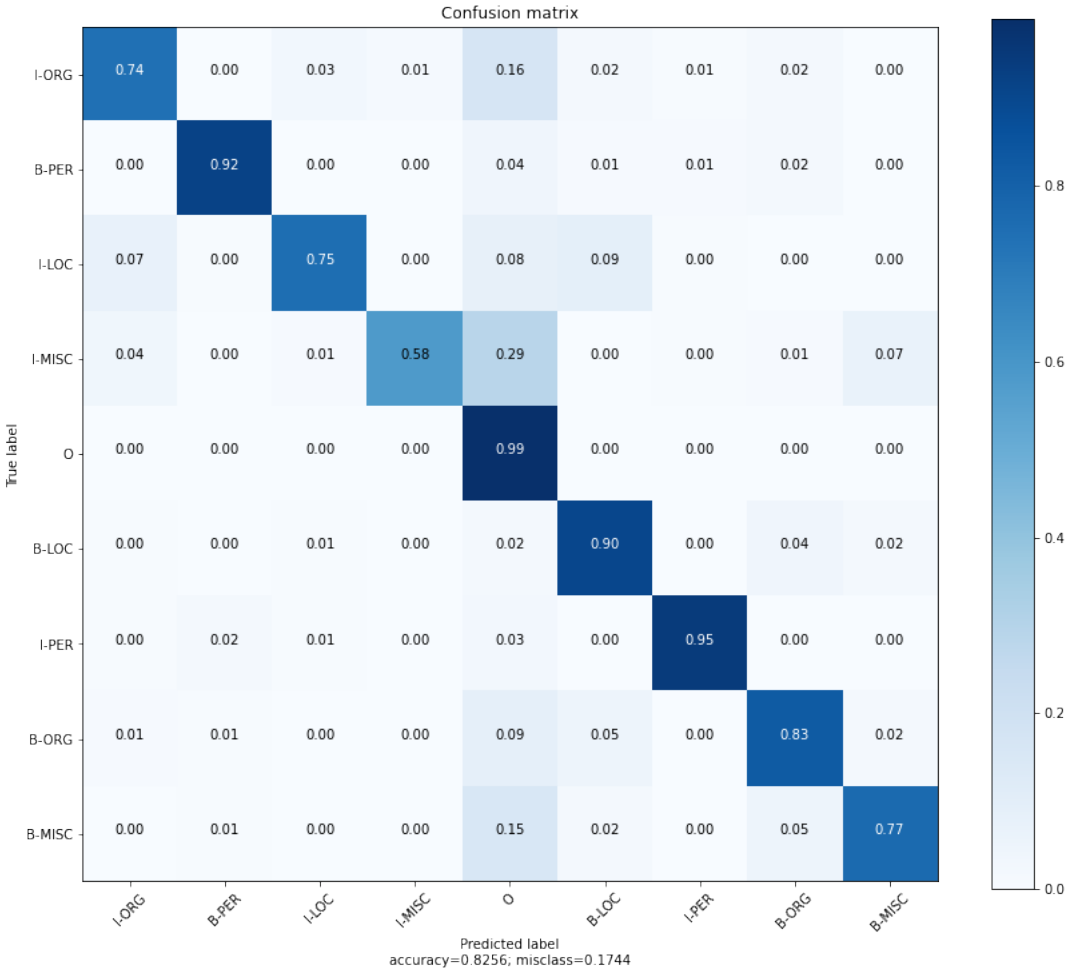


Figura 19 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 50 Épocas.

Figura 20 – Matriz de Confusão do Modelo de LSTM Pre-treinado com 100 Épocas.

APÊNDICE C – Valores das Precisões nas k posições

Tabela 16 – Valores das Precisões nas k_esimas Posições Utilizados no Gráfico de Precisão.

K	Precisão da Oclusão	Precisão Aleatória
1	65,52%	06,21%
2	43,45%	06,47%
3	32,99%	06,92%
4	26,55%	07,08%
5	21,86%	07,14%

APÊNDICE D – Palavras Anotadas e Não Anotadas por Posição no Rank

Tabela 17 – Quantidade de Palavras Anotadas e Não Anotadas por Posição no Rank.

Posição	Quantidade Anotada	Quantidade Não Anotada
1	190	659
2	62	787
3	35	814
4	21	828
5	9	840

Tabela 18 – Porcentagem de Palavras Anotadas e Não Anotadas por Posição no Rank.

Posição	Quantidade Anotada	Quantidade Não Anotada
1	59,93%	16,78%
2	19,56%	20,04%
3	11,04%	20,72%
4	06,62%	21,08%
5	02,84%	21,38%

APÊNDICE E – Gráficos da Palavras Relevantes por Rank Separados por Modelo.

Figura 21 – Distribuição dos Valores do Modelo 1 - Ranks 2, 3, 4, 5.

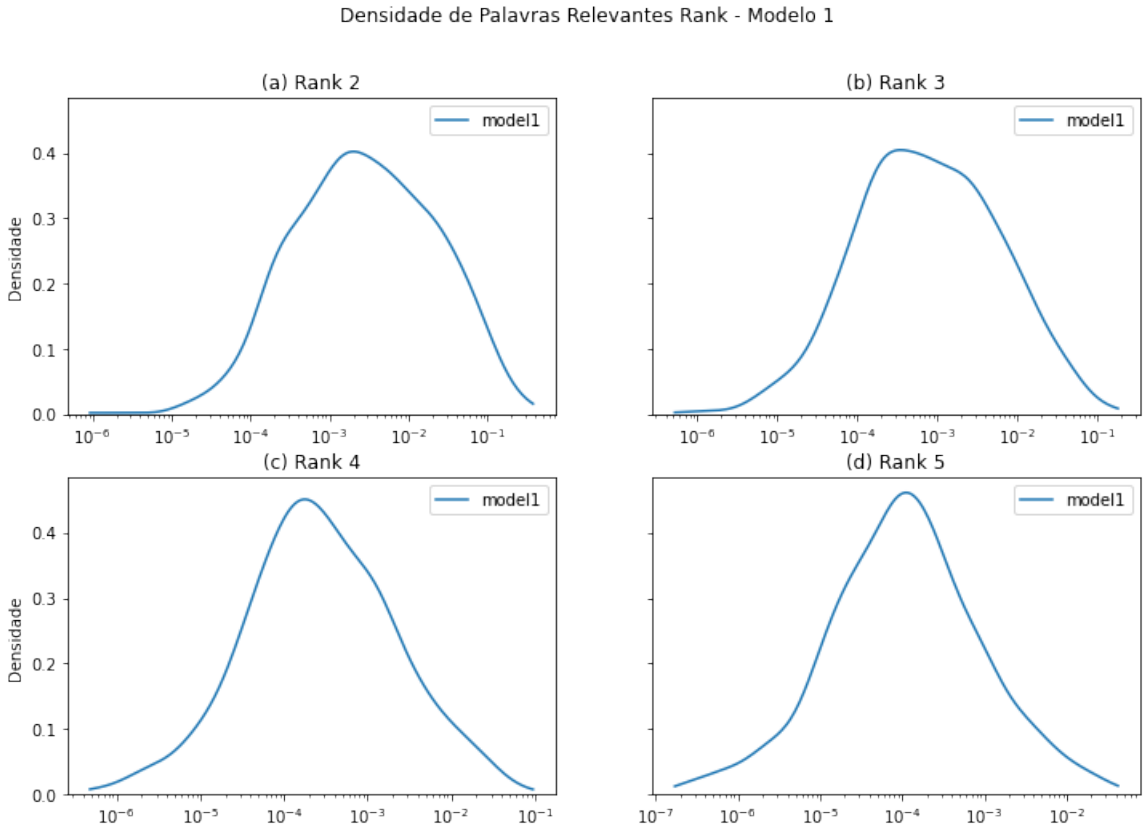


Figura 22 – Distribuição dos Valores do Modelo 2 - Ranks 2, 3, 4, 5.

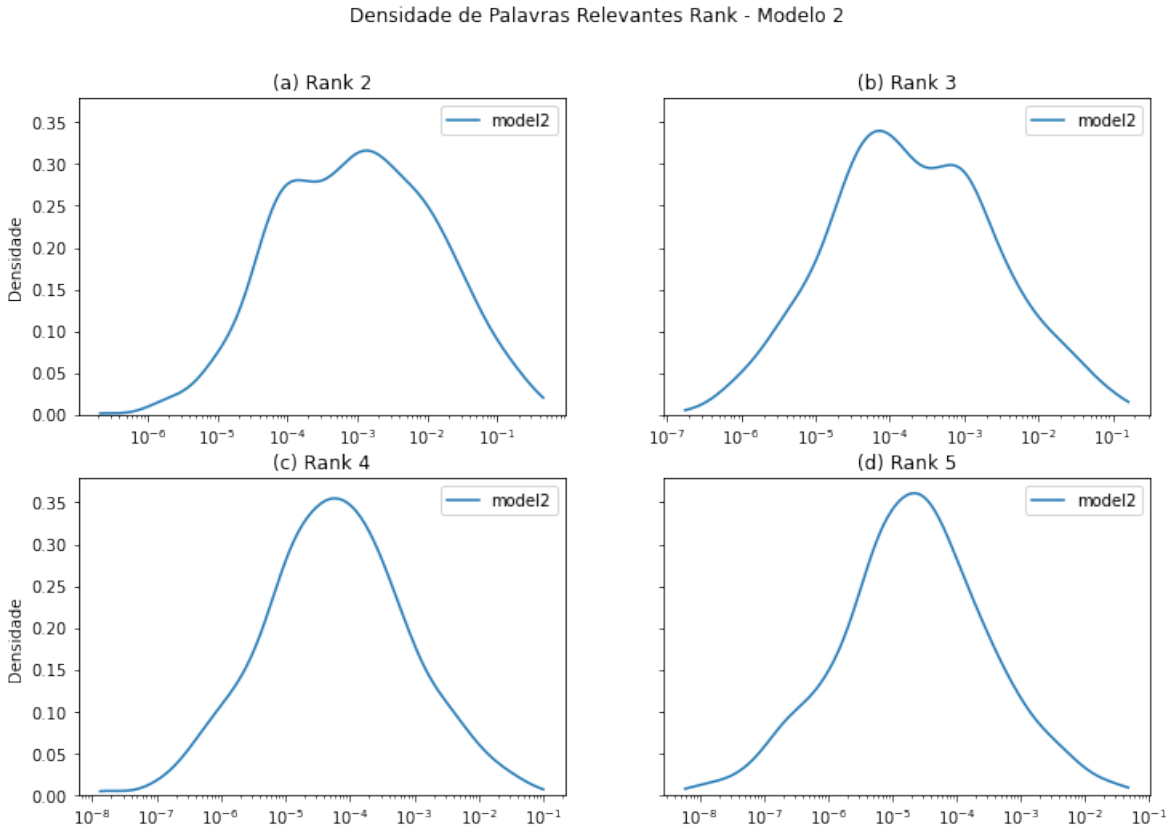


Figura 23 – Distribuição dos Valores do Modelo 3 - Ranks 2, 3, 4, 5.

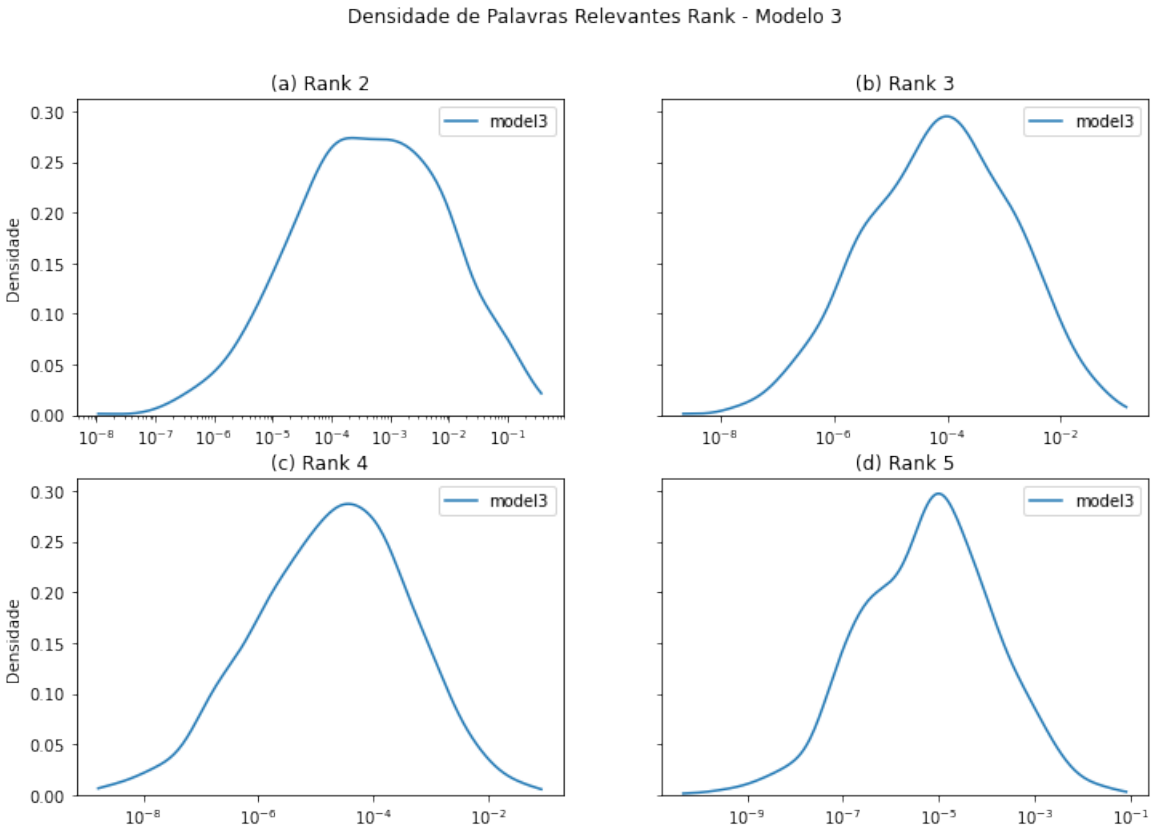


Figura 24 – Distribuição dos Valores do Modelo 4 - Ranks 2, 3, 4, 5.

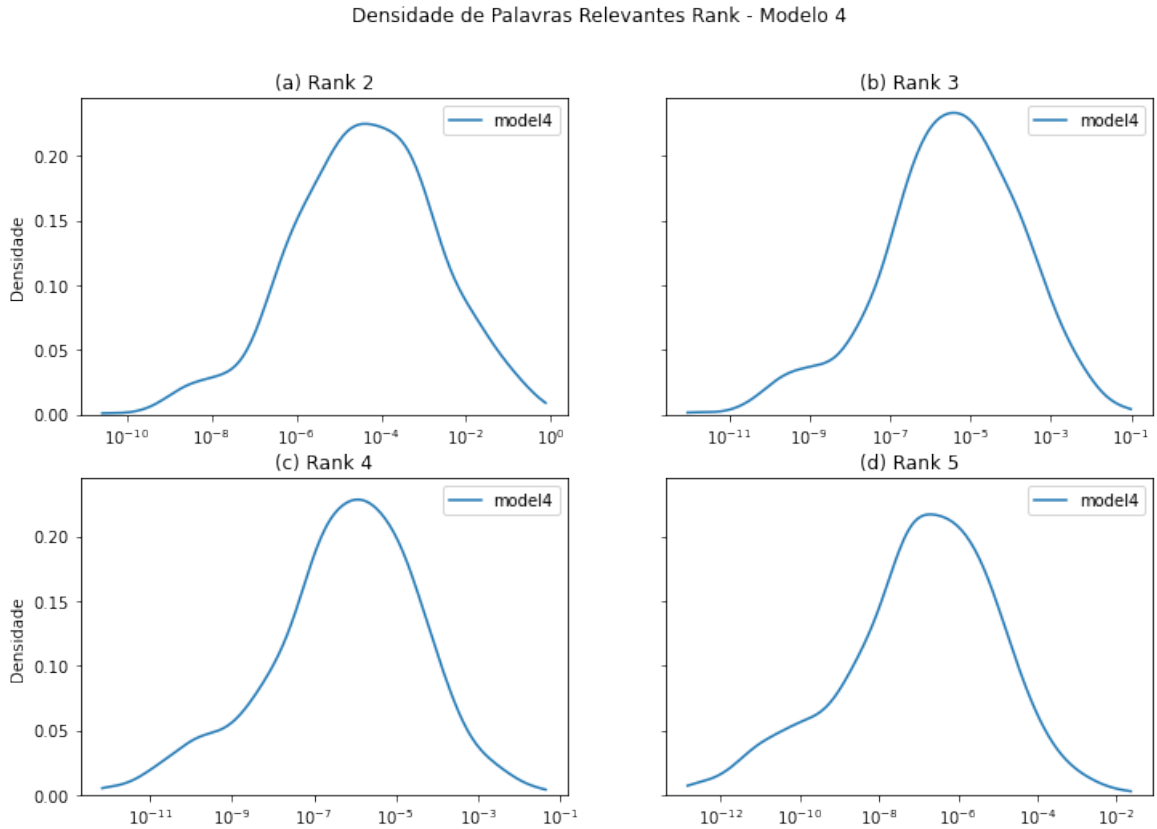
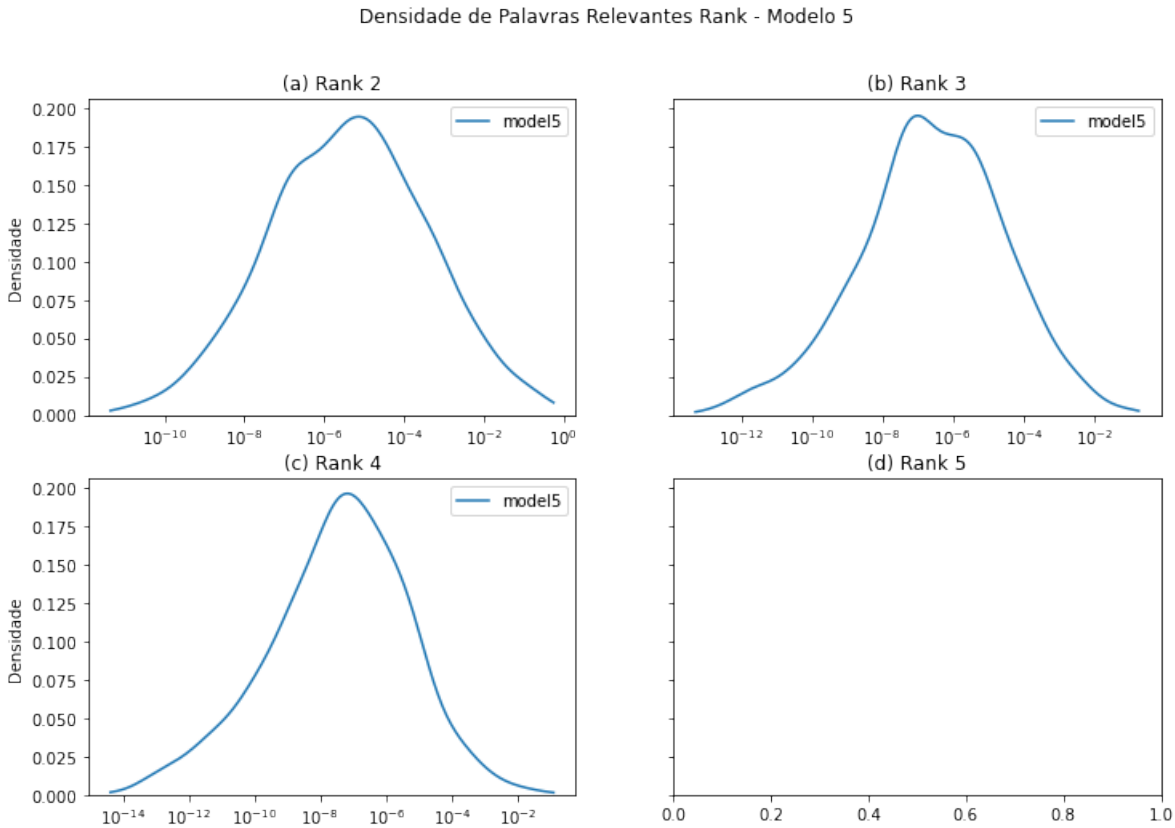



Figura 25 – Distribuição dos Valores do Modelo 5 - Ranks 2, 3, 4, 5.