INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL APLICADA À CLASSIFICAÇÃO DE DESINFORMAÇÃO

Pedro Lamkowski dos Santos
Orientador: João Paulo Papa
Coorientador: Gustavo Henrique Rosa

Agenda

Introdução

Interpretabilidade e explicabilidade

LRP e SS3

Desenvolvimento

Resultados

Conclusão

Introdução

Notícias falsas e desinformação atualmente

- Internet como meio de disseminação.
- *Fake news* - notícias falsas ou arquitetadas.
  - Afetam a opinião pública, eventos e eleições.
  - Rápida dispersão, especialmente em redes sociais.
- Jornalistas atualmente optam pelo desuso do termo *fake news*.

Causas

- Facilidade de produção e monetização;
- Barreira de entrada facilitada;
- Desconfiança nos meios de comunicação;
- Polarização política.

Classificação de notícias falsas

- Classificação de texto a partir de características de documentos.
- Tarefa do campo de Processamento de Linguagem Natural.
- O impacto social da desinformação torna a tarefa relevante.
- Bons resultados de classificação de modelos utilizando redes neurais profundas (*deep learning)*.
- Tarefa ainda desafiadora e abrangente.

A opacidade de modelos de deep learning

- Apresentam uma estrutura aninhada não linear.
- Não apresentam informações sobre a tomada de decisão.
- Muitos considerados abordagens de “caixa-preta”.
- Inconveniência em aplicações que necessitam de transparência.

Inteligência Artificial Explicável (XAI)

Técnicas que buscam o entendimento de modelos de Inteligência Artificial e Aprendizado de Máquina e explicar predições individuais.

Objetivo

Avaliar duas técnicas de XAI na tarefa de classificação de desinformação utilizando dos modelos de aprendizado de máquina, sendo um inerentemente explicável e outro arquitetado como uma rede neural.

Interpretabilidade e explicabilidade

O que é uma explicação?

- Precisamos saber o que queremos perguntar a um modelo.
- A explicação depende da pergunta.
- A explicação pode ser avaliada quanto sua:
  - Interpretabilidade;
  - Completude.

Categorias de XAI

Podemos dividir XAI em duas categorias:

- Inerente;
- *post-hoc*.

Motivações de indivíduos de interesses chaves na explicabilidade.

Indivíduo de Interesse	**Motivação**
Cientista de dados	Entender o modelo; retirar bugs; melhorar a performance.
Empresário	Entender o modelo; avaliar validade para o propósito; aceitar o uso.
Analista de risco de modelo	Desafiar o modelo; assegurar da robustez; aprovar o uso.
Regulador	Verificar impacto nos consumidores; verificar a confiabilidade.
Consumidor	Entender o impacto do modelo em sua vida e tomada de ações.

Perspectivas de explicabilidade

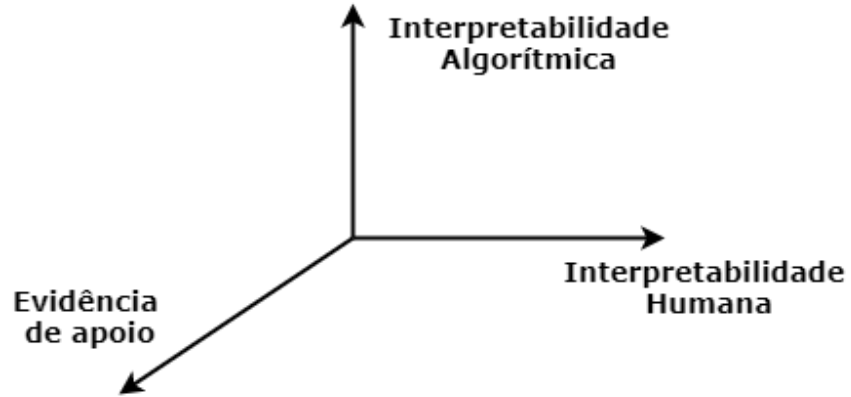
- Transparência;
- Critério de avaliação;
- Tipo de explicação.

Vantagens e desvantagens dos tipos de explicações.

Tipo de Explicação	Vantagens	Desvantagens
Explicações locais	Explica o comportamento do modelo em uma área local de interesse	Explicações não generalizam em uma escala global. Perturbações pequenas podem resultar em explicações diferentes. A definição de localidade é complexa. Algumas abordagens são instáveis.
Exemplificações	Exemplos representativos que proveem intuições sobre o funcionamento interno do modelo. Alguns algoritmos revelam os dados de treinamento que levam o modelo a suas predições.	Exemplos precisam de inspeção humana. Não ressaltam quais partes do exemplo influenciam o modelo.
Relevância das Características	Operam em cada instância do modelo, verificando a importância de cada característica de entrada na decisão do modelo. Várias abordagens propostas são acompanhadas de garantias teóricas.	São sensíveis em casos que as características são correlatas. Em vários casos, as soluções exatas são aproximadas, levando a efeitos indesejados, podendo afetar o resultado.
Simplificações	Modelos substitutos que explicam os incertos. Explicações resultantes, como regras, são facilmente entendíveis.	Modelos substitutos podem não estimar os modelos originais apropriadamente, além de terem suas próprias limitações.
Visualizações	De fácil comunicação com leigos. A maioria das abordagens são intuitivas e descomplicadas de implementar.	Existe um limite de quantas características podem ser representadas de uma vez. Humanos precisam inspecionar gráficos resultantes para produzir as explicações.

Classificação de desinformação explicável

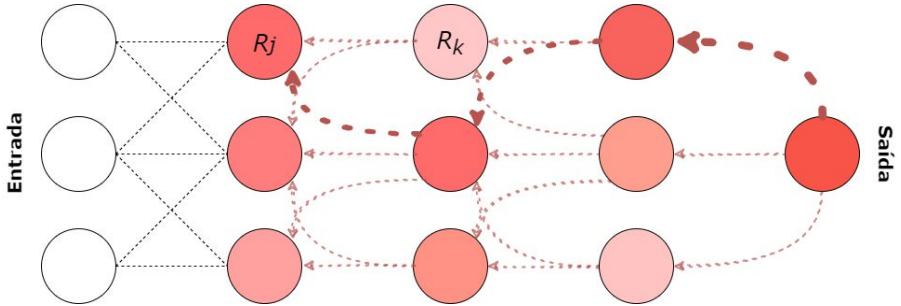
Sistemas de detecção de notícias falsas auxiliados de metadados são mais robustos.



LRP e SS3

Layer-wise relevance propagation

- Técnica de explicabilidade de redes neurais.
- Propagação do valor de predição no sentido contrário da rede.



$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$$

Regras do LRP

A utilização das regras depende da tarefa em que o LRP será aplicado.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$$

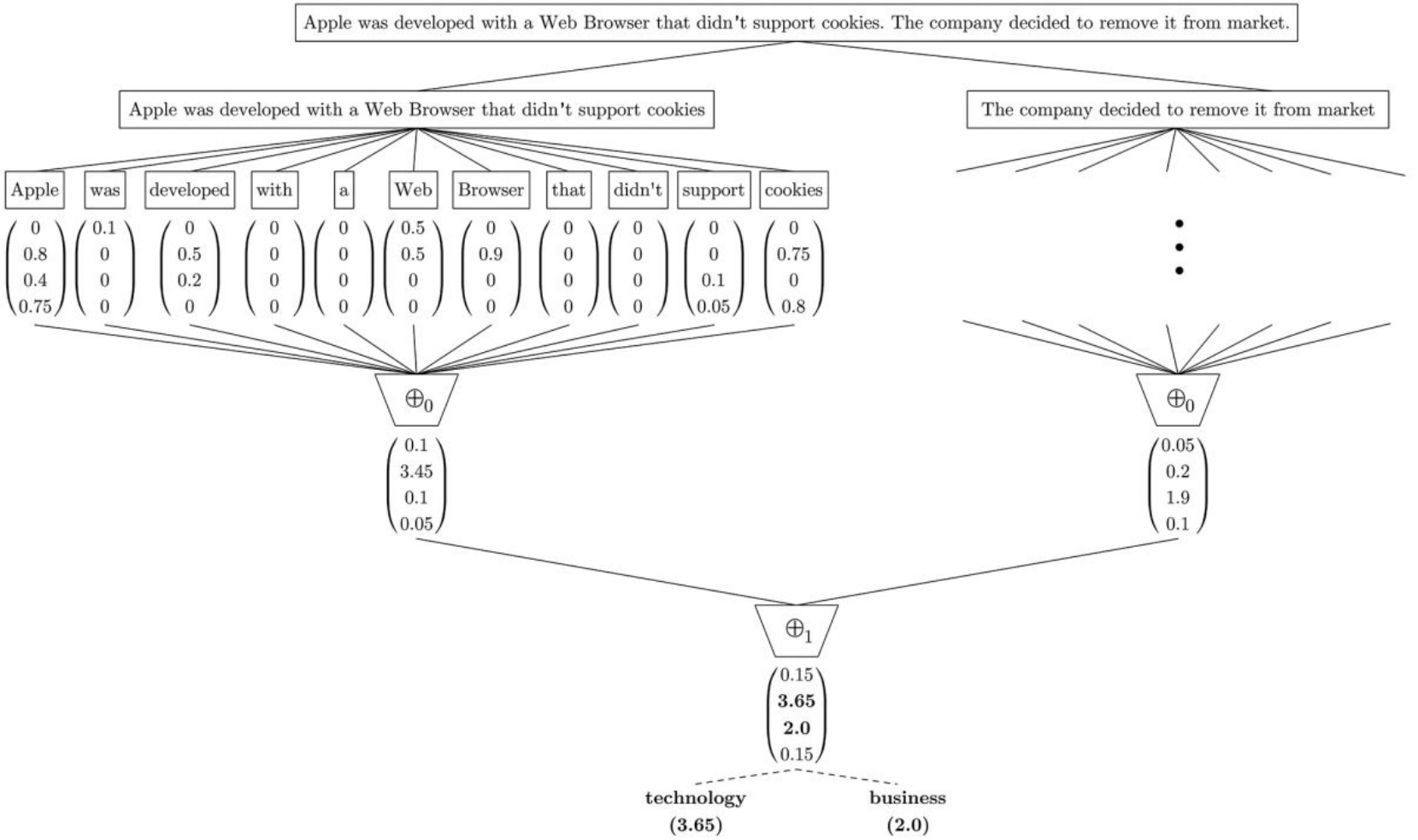
$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k$$

$$R_j = \sum_k (\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-}) R_k$$

Sequential S3 - Smoothness, Significance and Sanction

- Modelo de classificação de texto arquitetado para ser explicável.
- Cálculo do valor global de palavras, frases e documentos para classificação.
- Possibilidade de utilização de n-gramas no treinamento.

$$gv(w,c) = lv_{\sigma}(w,c) \cdot sg_{\lambda}(w,c) \cdot sn_{\rho}(w,c)$$


Apple was developed with a Web Browser that didn't support cookies. The company decided to remove it from market.
Apple was developed with a Web Browser that didn't support cookies
The company decided to remove it from market
Apple
was
developed
with
a
Web
Browser
that
didn't
support
cookies
$\oplus_0$
$\oplus_0$
$\oplus_1$
technology
(3.65)
business
(2.0)

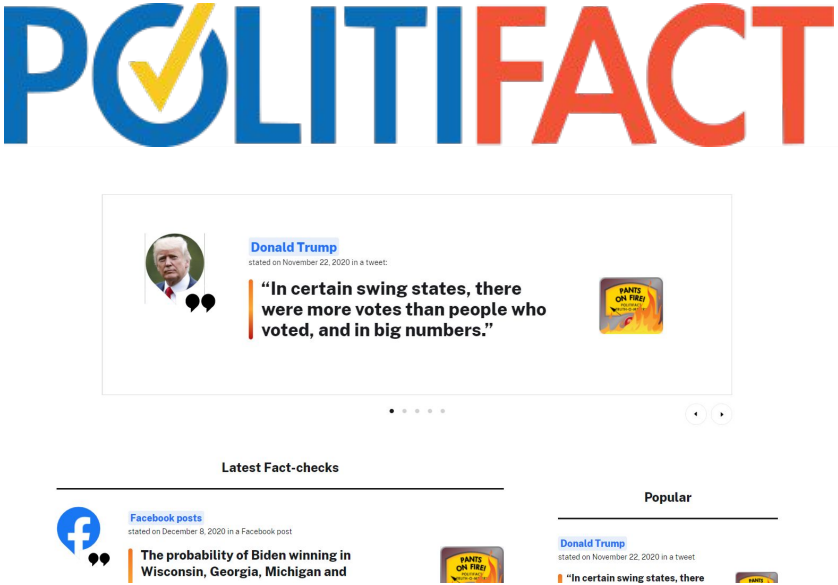
Desenvolvimento

Ferramentas

- Bibliotecas Python;
- PySS3;
- iNNvestigate;
- NLTK.

Conjuntos de dados

- **LIAR**
- Fake News Inference Dataset (FNID)
  - **FNID-LIAR**
  - **FNID-FNN** (FakeNewsNet)



Divisão dos conjuntos de dados

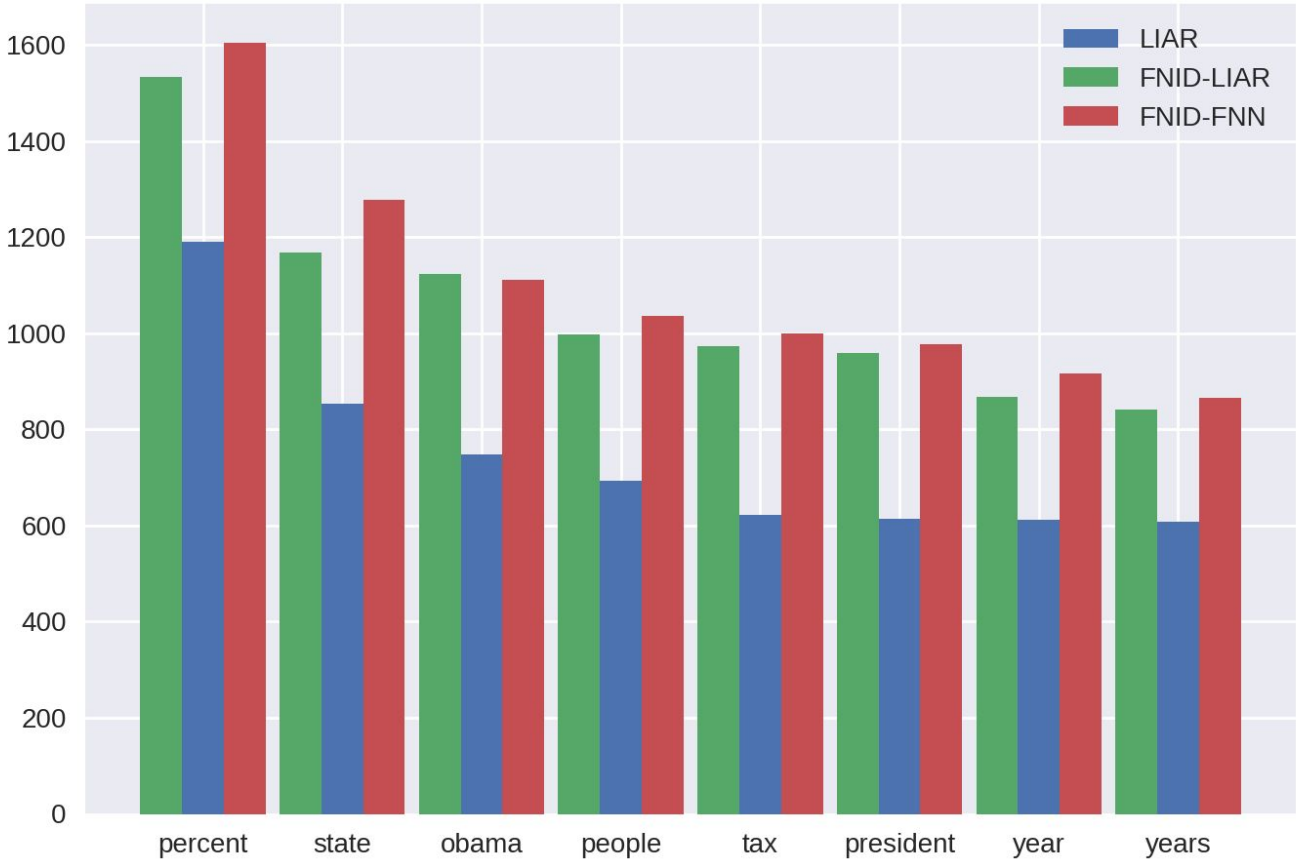
LIAR

Conjunto de treinamento	10269
Conjunto de validação	1284
Conjunto de teste	1283
Tamanho médio de *token*	17.9

FNID

FNID-LIAR	
Treinamento	15052
Validação	1265
Teste	1266
FNID-FNN	
Treinamento	15212
Validação	1058
Teste	1054

Frequência de tokens nos conjuntos de dados.



Tratamento dos dados

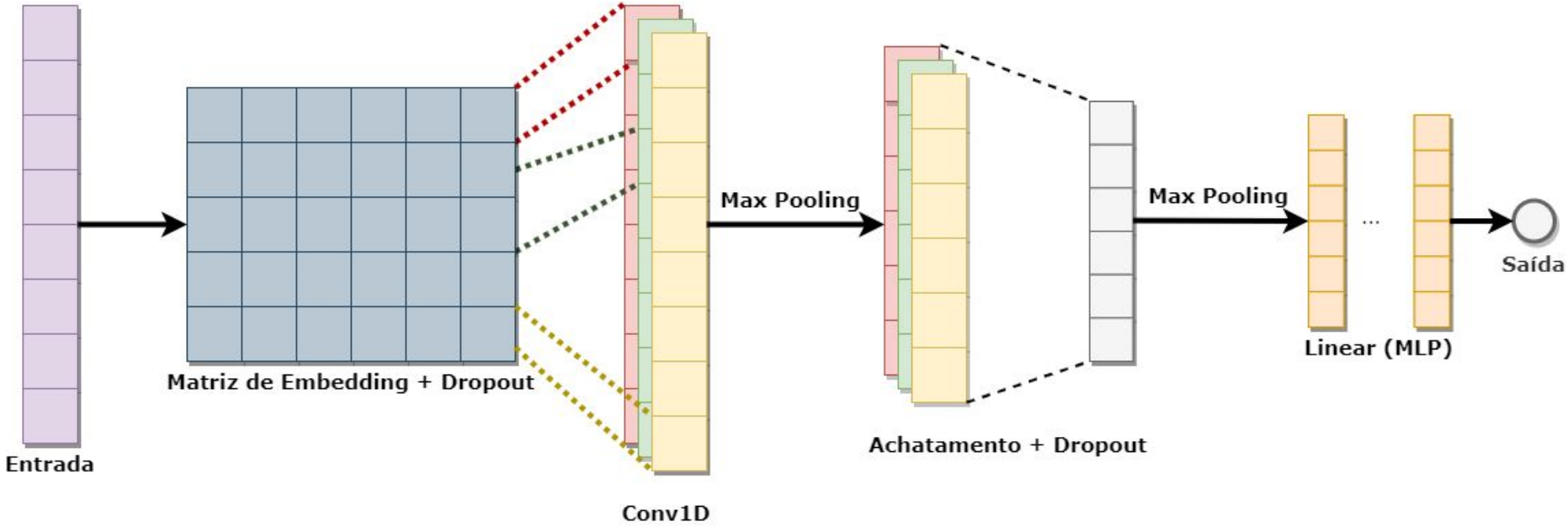
Etapa de pré-processamento

- Remoção de caracteres dispensáveis;
- Remoção de palavras vazias;
- *Stemming* das palavras;
- Criação de *loaders*;
- Rótulos das notícias organizados em verdadeiro e falso.

Modelos

- SS3.
- Rede neural convolucional (CNN) 1D.

Arquitetura da rede convolucional.

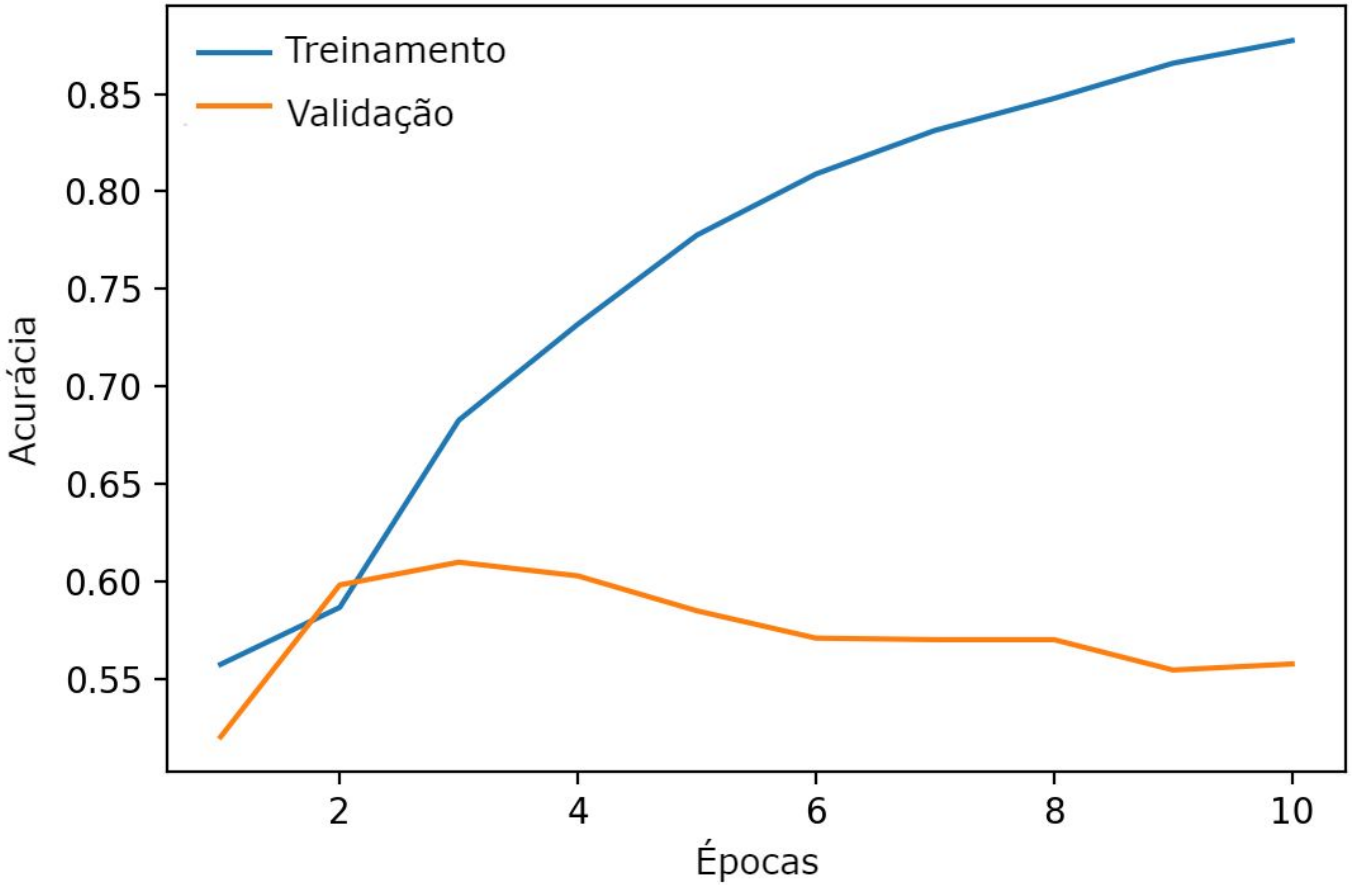


Treinamento

Ambos modelos treinados com os conjuntos de treinamento de cada conjunto. Primeira análise utilizando conjunto de validação.

- SS3: utilização de n-gramas (n=2 e n=3) não trouxe melhorias.
- CNN: Modelo treinado por 10 épocas, modelo sofreu *overfitting*.

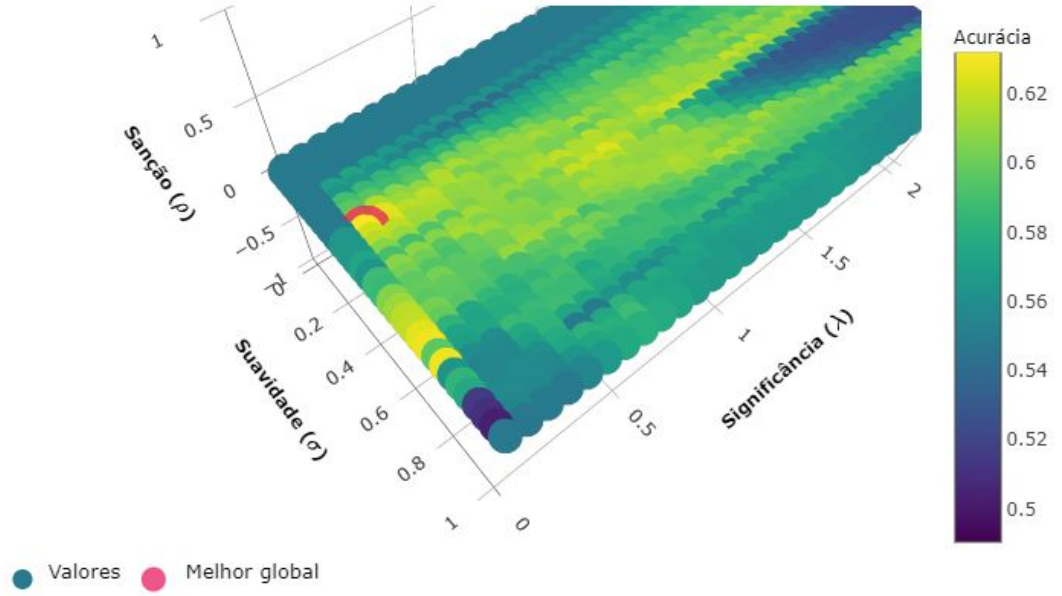
Comparação da variação da acurácia de treinamento e validação no conjunto LIAR indicando *overfitting*.



Acurácia no conjunto de validação sem otimização.

Otimização

- Busca em grade para seleção dos melhores hiper-parâmetros.
- Diminuição do número de épocas no treinamento.



Resultados

Avaliação dos modelos

Análise utilizando os conjuntos de teste, com os parâmetros encontrados na otimização.

SS3

	Acurácia	Precisão	Sensitividade	Pontuação F1
LIAR	0.6133	0.6035	0.5953	0.5941
FNID-LIAR	0.5987	0.5976	0.5981	0.5900
FNID-FNN	0.6395	0.7219	0.6910	0.6354

CNN

	Acurácia	Precisão	Sensitividade	Pontuação F1
LIAR	0.6196	0.5877	0.4363	0.4919
FNID-LIAR	0.5940	0.5394	0.5381	0.5295
FNID-FNN	0.6850	0.5134	0.6488	0.5203

Avaliação dos modelos

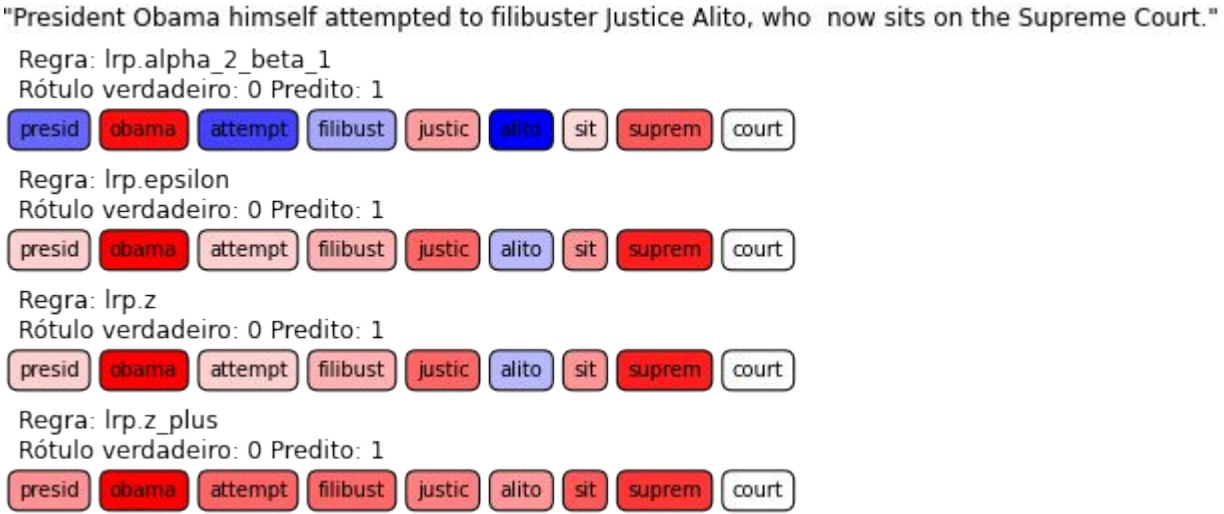
- Performance razoável, mas deixam a desejar comparado a arquiteturas mais complexas de redes neurais.
- Mesmo sendo mais simples, SS3 teve uma performance comparável ao CNN.
- O conjunto **FNID-FNN** teve melhor performance em ambos modelos:
  - SS3 Acurácia: 0.6395
  - CNN Acurácia: 0.6850

Visualizando explicações

- Utilizando os modelos no conjunto FNID-FNN.
- Regras de LRP utilizadas:
  - LRP-$\alpha\beta$;
  - LRP-ϵ;
  - LRP-z;
  - LRP-z+.

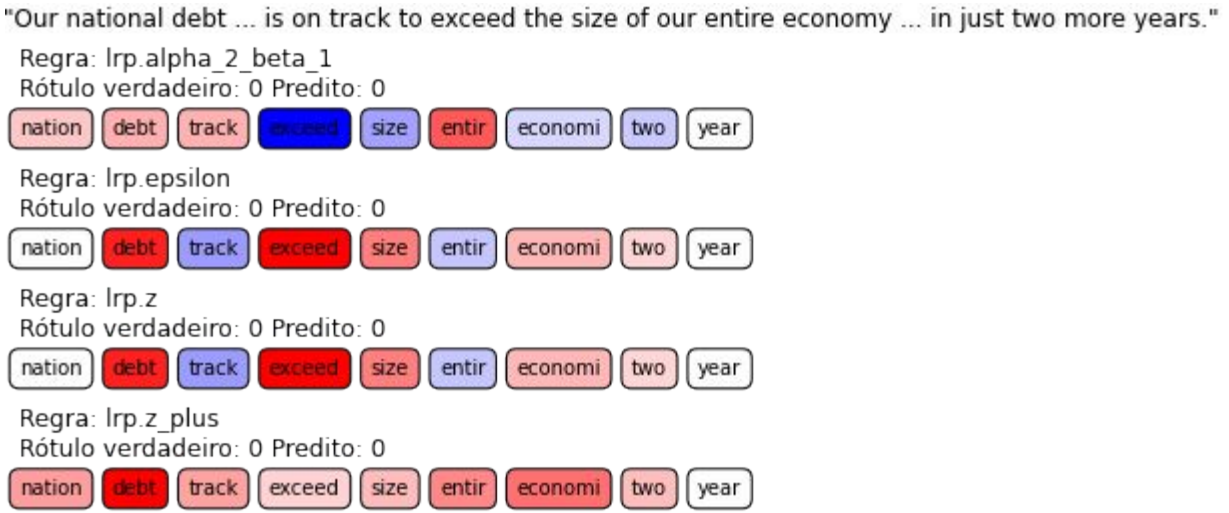
Visualizando explicações (LRP)

- Verificados vieses no conjunto de dados.



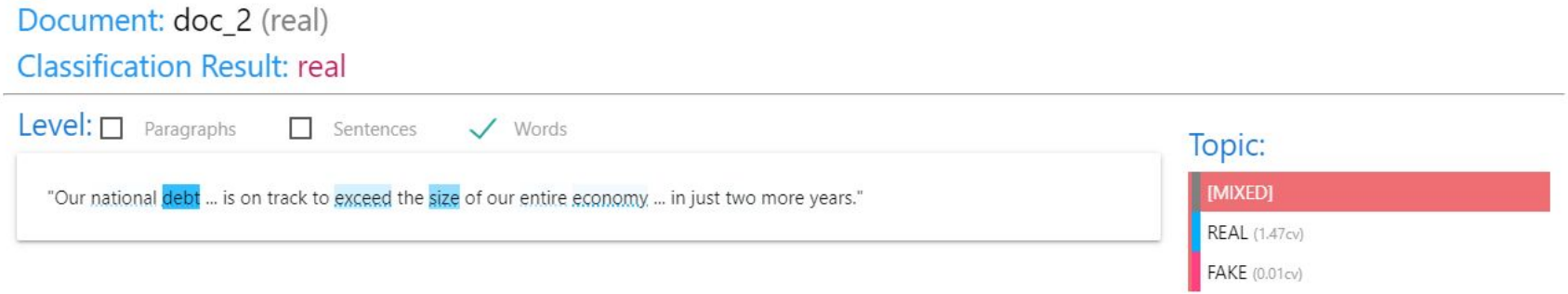
Visualizando explicações (LRP)

- A escolha da regra é importante para a interpretabilidade.
- Regras LRP-ϵ e LRP-z atribuem uma pontuação ruidosa.



Visualizando explicações (Live_test)

As explicações do SS3 corroboram para as mesmas ideias do LRP.



Conclusão

Conclusão

- Utilização de modelos de Aprendizado de Máquina simples para uma tarefa desafiadora.
- Estudo e abordagem de conceitos de explicabilidade e interpretabilidade em IA.
- Estudo de um modelo recente para classificação de texto.
- Emprego de tecnologias e ferramentas de aprendizado de máquina não vistas na graduação.

Trabalhos futuros

- Analisar modelos de classificação de desinformação mais robustos.
- Utilizar XAI em outras áreas e tarefas.

Referências

- ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. Journal of economic perspectives, v. 31, n. 2, p. 211–36, 2017.
- BELLE, V.; PAPANTONIS, I. Principles and practice of explainable machine learning. arXiv preprint arXiv:2009.11698, 2020.
- LAZER, D. M.; BAUM, M. A.; BENKLER, Y.; BERINSKY, A. J.; GREENHILL, K. M.; MENCZER, F.; METZGER, M. J.; NYHAN, B.; PENNYCOOK, G.; ROTHSCHILD, D. et al. The science of fake news. Science, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018.
- REIS, J. C. S.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Explainable machine learning for fake news detection. In: Proceedings of the 10th ACM Conference on Web Science. New York, NY, USA: Association for Computing Machinery, 2019. p. 17–26.

MUITO OBRIGADO!