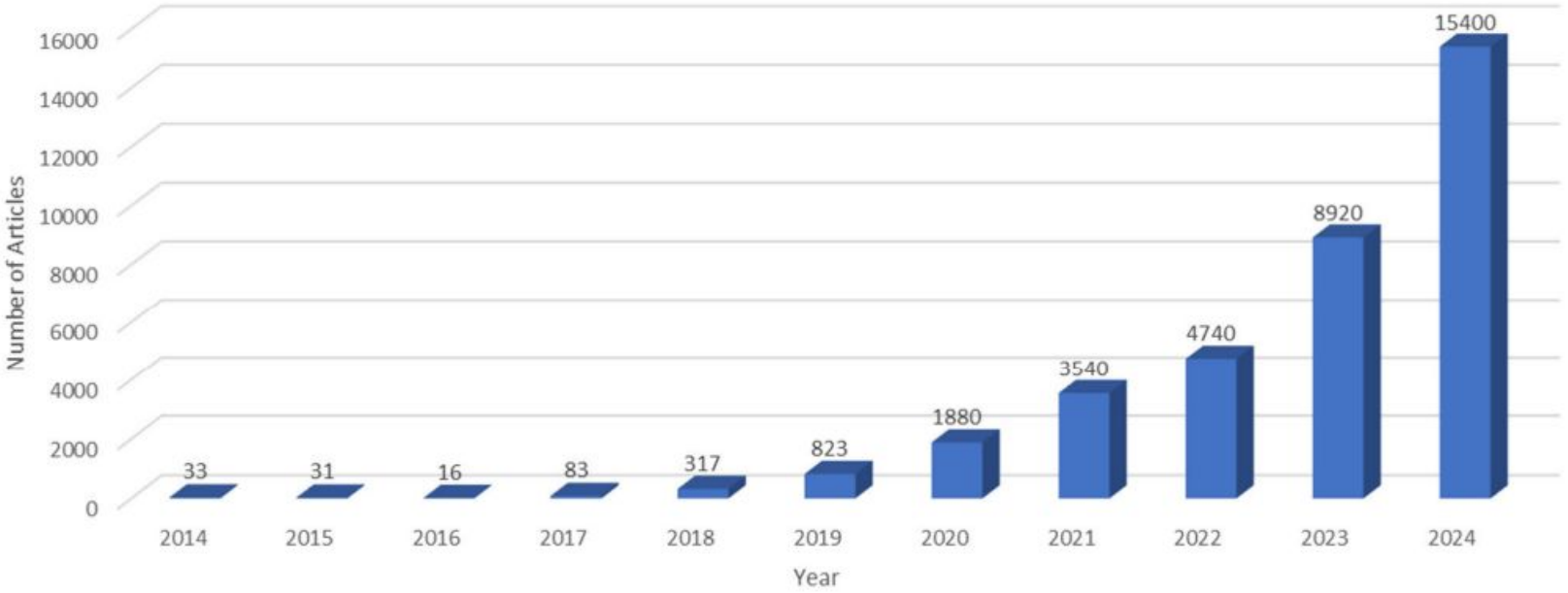
Utilização de XAI para explicação de diferentes encoders

Cesar Augusto Pereira Vial
Paulo Sérgio Ávila Júnior

Introdução

O advento do XAI nos últimos anos tem se dado pela:

- Popularização do uso de métodos “Caixa-Preta”, como as LLMs, devido a sua alta acurácia;
- Potencial do uso desses métodos em áreas críticas, como saúde e finanças;
- Promessa de mais *Transparência* e *Segurança* para o uso de IA.

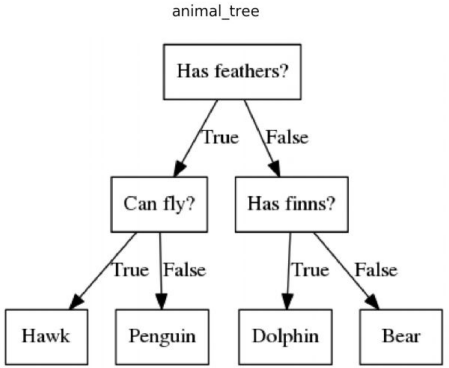


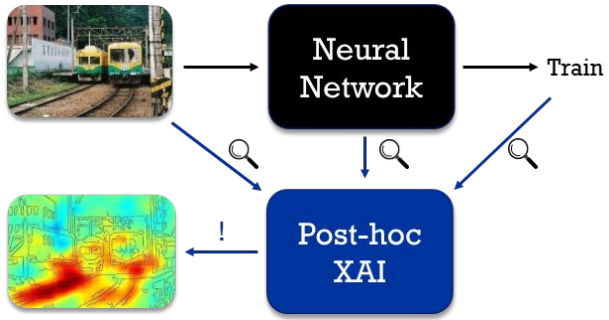
Número de artigos por ano com XAI como tema

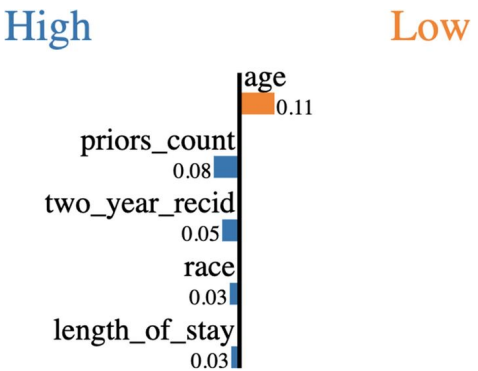
eXplainable Artificial Intelligence

Métodos mais populares:

- Modelos de Interpretabilidade Intrínseca (modelos que podem ser interpretados por si só);
- Métodos post-hoc (buscam explicar os resultados obtidos pelos modelos), comumente utilizados com
- Ferramentas de Visualização (auxiliam na visualização do funcionamento do modelo).







Trabalhos Recentes

Os trabalhos recentes em XAI buscam, entre outras linhas de pesquisa:

- Aumentar a *Transparência* e *Interpretabilidade* dos modelos de IA mais populares, como as LLMs;
- Possibilitar o uso de modelos de IA em áreas críticas, como saúde e finanças;
- Encontrar métodos *model-agnostic* e padronizados de XAI;
- Desenvolver medidas de performance para as XAI, de modo a possibilitar a comparação entre técnicas diferentes.

Trabalhos Recentes - Saúde

Na área de saúde estão sendo utilizadas:

- As técnicas de SHAP e LIME para explicar predições médicas e diagnósticos;
- O GradCAM para explicações em análises de imagem;
- Os modelos de Árvores de Decisão e Redes Bayesianas, devido a sua interpretabilidade intrínsica;
- Mecanismos de Atenção e Mapas de Saliência, para enfatizar regiões críticas em diagnósticos por imagem;
- Métricas de avaliação de XAIs, para fornecer uma base na decisão de qual modelo utilizar tendo em vista a criticidade e a velocidade necessária para determinada tarefa.

Trabalhos Recentes - Finanças

Na área de finanças o XAI, principalmente a partir das técnicas SHAP e LIME, está sendo utilizado para:

- Detecção de Fraude não-invasiva;
- Análise de risco de crédito;
- Predição de mercado.

Trabalhos Recentes - Outras áreas

- **Educação:** Geração de explicações automáticas para ensino; Análise de dados vocacionais.
- **Cibersegurança:** Detecção de *malware* e inspeção de tráfego de rede.
- **Lei:** Interpretação de casos; análise de contratos; auxílio a decisões judiciárias.
- **Agricultura:** Sistemas de recomendação de plantio; análise de fatores que influenciam na qualidade dos produtos.

Proposta

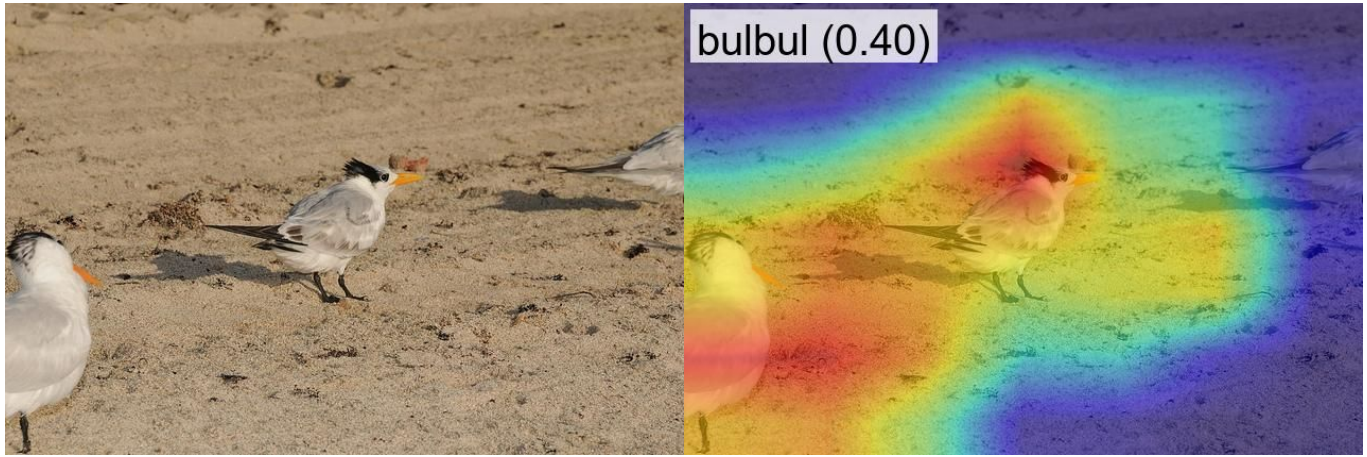
- Relacionar Descrições e Imagens, utilizando o método de relacionamento de *embeddings* de textos e imagens;
    - Artigo base: *Maniparambil et al. 2024: Do vision and language encoders represent the world similarly?*
- Agrupar imagens e descrições, e relacionar a explicabilidade de ambos, utilizando algoritmos de XAI com os *embeddings* criados e as relações encontradas.

Proposta

- Utilização do *COCO Dataset*, para imagens e descrições;
- Para a criação dos *embeddings* de texto foi utilizado o modelo *RoBERTa*;

A bird sits in the medium brown colored sand.
A small white bird standing on top of a dirt field.

- Para a criação dos *embeddings* de imagem foi utilizado o modelo *ResNet50*;

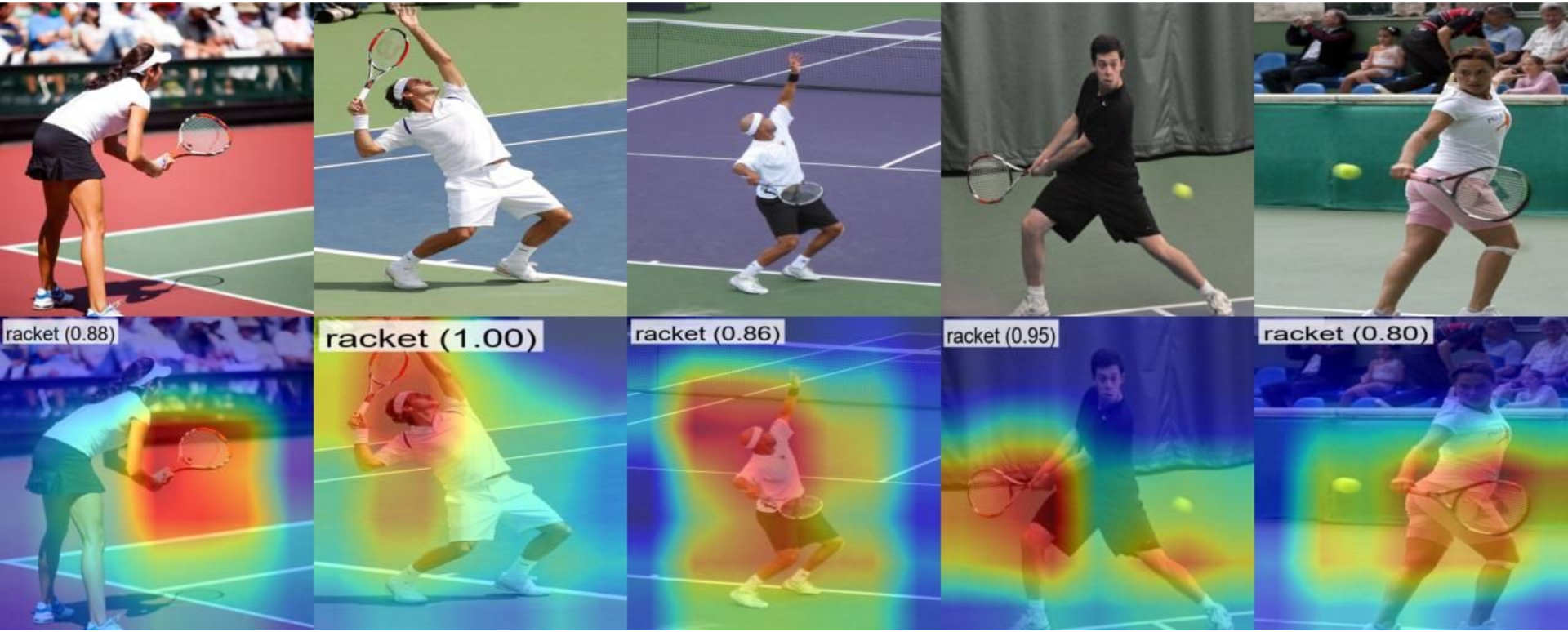


Experimentos

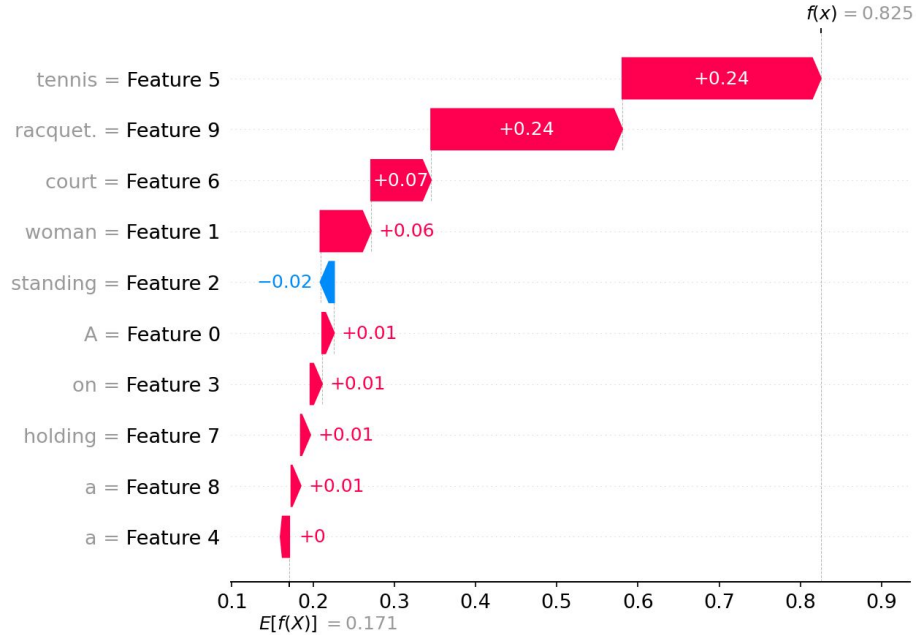
- Para a explicabilidade do modelo de texto foi utilizado o algoritmo SHAP, utilizando as palavras como *features*;
- Para a explicabilidade do modelo de imagens foi utilizado o algoritmo Grad-CAM, diretamente em cima da classe com maior probabilidade de acordo com o *ResNet50*.

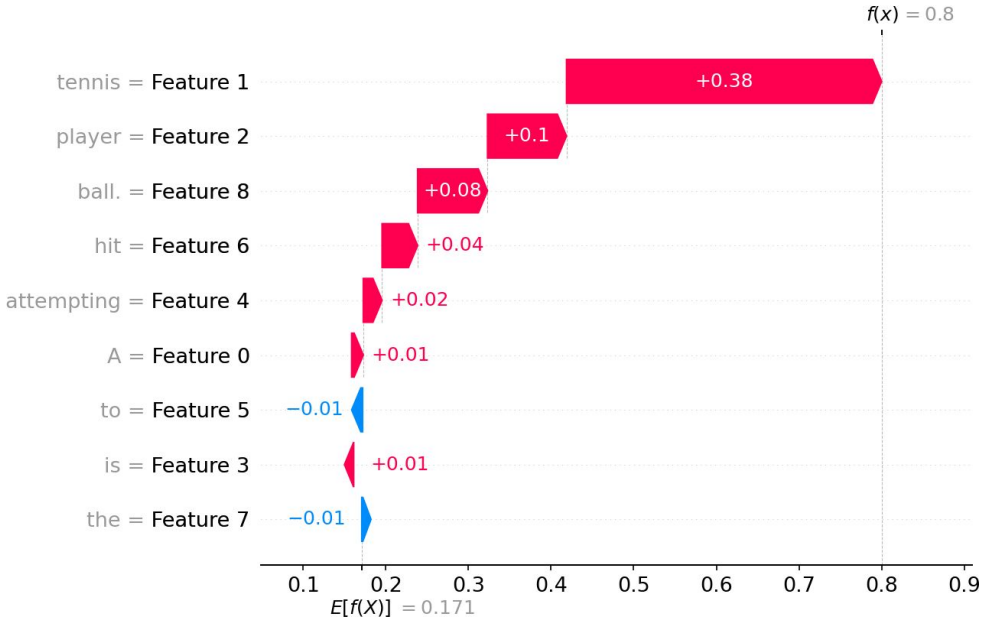
Resultados

A woman standing on a tennis court holding a racquet. / A tennis player is attempting to hit the ball.



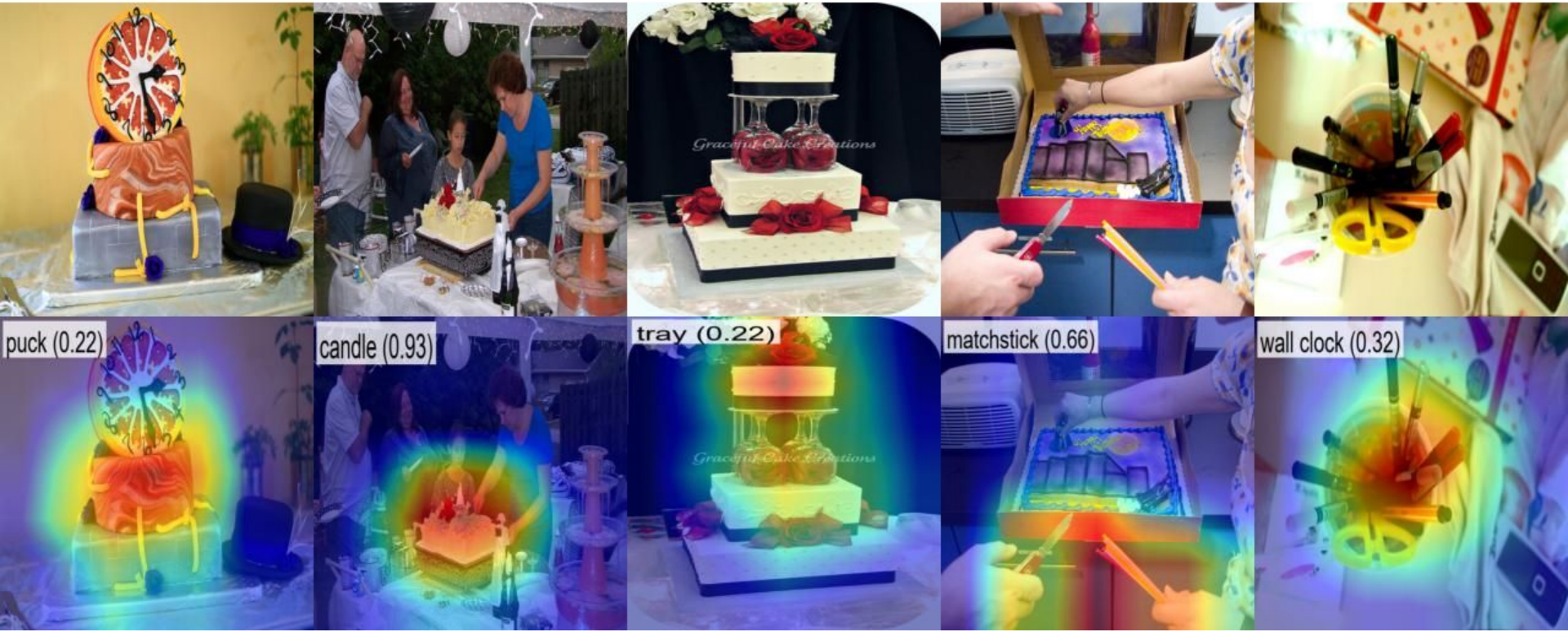
Resultados


f(x) = 0.825
tennis = Feature 5
racquet. = Feature 9
court = Feature 6
woman = Feature 1
standing = Feature 2
A = Feature 0
on = Feature 3
holding = Feature 7
a = Feature 8
a = Feature 4
+0.24
+0.24
+0.07
+0.06
−0.02
+0.01
+0.01
+0.01
+0.01
+0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
E[f(X)] = 0.171

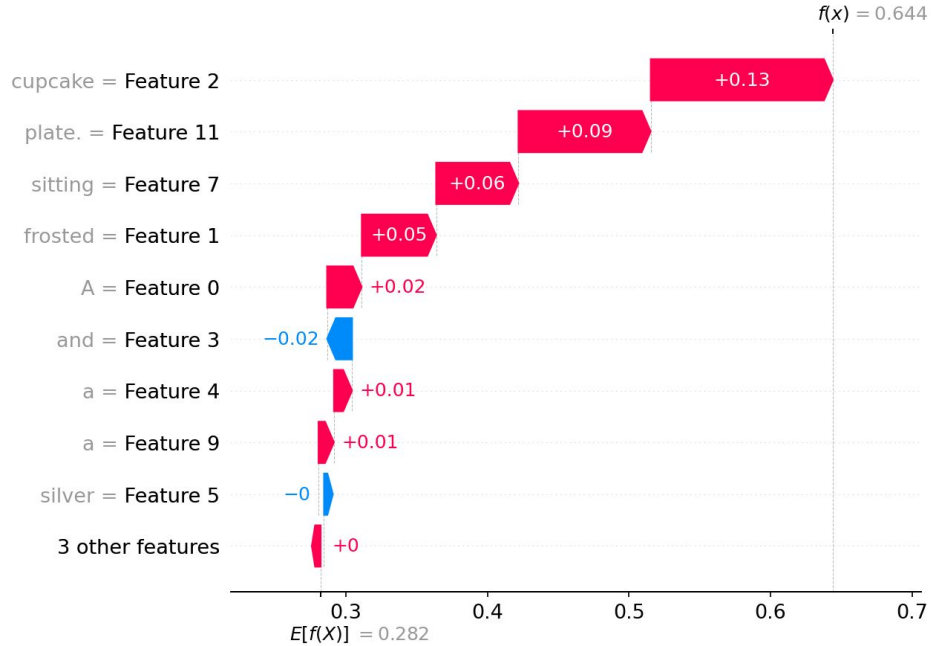

f(x) = 0.8
tennis = Feature 1
player = Feature 2
ball. = Feature 8
hit = Feature 6
attempting = Feature 4
A = Feature 0
to = Feature 5
is = Feature 3
the = Feature 7
+0.38
+0.1
+0.08
+0.04
+0.02
+0.01
−0.01
+0.01
−0.01
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
E[f(X)] = 0.171

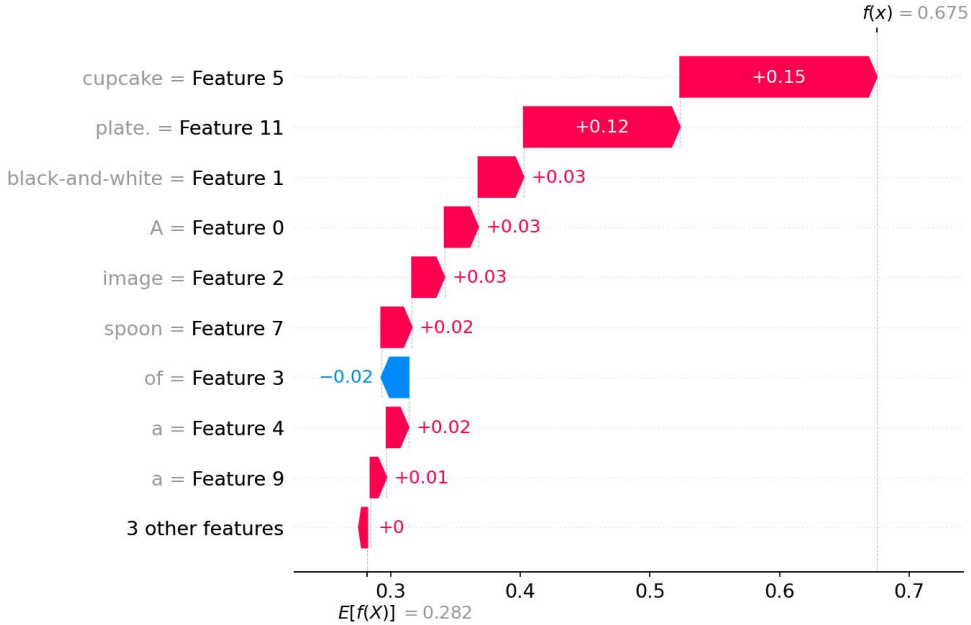
Resultados

A frosted cupcake and a silver spoon sitting on a white plate. / A black-and-white image of a cupcake and spoon on a small plate.



Resultados


f(x) = 0.644
cupcake = Feature 2
plate. = Feature 11
sitting = Feature 7
frosted = Feature 1
A = Feature 0
and = Feature 3
a = Feature 4
a = Feature 9
silver = Feature 5
3 other features
+0.13
+0.09
+0.06
+0.05
+0.02
−0.02
+0.01
+0.01
−0
+0
0.3
0.4
0.5
0.6
0.7
E[f(X)] = 0.282


f(x) = 0.675
cupcake = Feature 5
plate. = Feature 11
black-and-white = Feature 1
A = Feature 0
image = Feature 2
spoon = Feature 7
of = Feature 3
a = Feature 4
a = Feature 9
3 other features
+0.15
+0.12
+0.03
+0.03
+0.03
+0.02
−0.02
+0.02
+0.01
+0
0.3
0.4
0.5
0.6
0.7
E[f(X)] = 0.282

Conclusão

- O modelo do *ResNet50* possui alguns focos inesperados;
- Para a explicabilidade do modelo *RoBERTa*, é verificado um claro foco em substantivos e adjetivos, o que facilita a visualização das palavras importantes nas imagens;
- Poucas palavras colaboram de forma negativa para explicar seu resultado. Essas palavras são, em sua maioria, artigos e preposições.

Trabalhos Futuros - Experimentos

- Utilização de diferentes modelos para a criação dos *embeddings*;
- Utilização de diferentes algoritmos de XAI;
- Para a análise do *Grad-CAM*, utilizar a descrição relacionada, e não o resultado direto do modelo utilizado nos *embedds*;
- Realizar a análise completa em modelos generativos ou de busca de *image-to-text* ou *text-to-image*.

Trabalhos Futuros

Os trabalhos futuros de XAI envolvem:

- Criação de *benchmarks* e métricas para avaliação da performance das técnicas de XAI;
- Redução do *overhead* computacional relacionado aos métodos de XAI mais populares;
- Integração de *guidelines* éticos aos *frameworks* de XAI;
- Mapeamento das áreas críticas que exigem o uso de XAI na indústria;
- Mapeamento de itens das leis gerais de proteção de dados (LGPD, GPDR) que se alinham com o uso de XAI;
- Proposição de *frameworks* para o uso de XAI por leigos.

Obrigado!