Classificação multimodal para detecção de produtos proibidos em uma plataforma *marketplace*

Alan da Silva Romualdo[1], **Livy Real**[2], **Helena de Medeiros Caseli**[1]

[1]Universidade Federal de São Carlos (UFSCar) – Departamento de Computação
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

[2]Digital Lab – americanas s.a.
São Paulo – SP – Brasil

alan.romualdo@b2wdigital.com, helenacaseli@ufscar.br

livy.coelho@b2wdigital.com

Abstract. *The multimodal learning aims to explore the characteristics of different modalities (text, image, audio) to generate computational models. In electronic commerce, due to the great variety of product features and the absence or inconsistency of information, the combination of information from different modes is quite adequate. This work presents some experiments caried out for the multimodal classification (text and image) of products (adult products) that cannot be sold in the marketplace of the partner company. In these experiments, neural networks were used to train uni and multimodal classifiers. The multimodal classifier achieved* 99% *of* $F1$ *against* 98% *for the textual model and* 94% *for the visual one.*

Resumo. *O aprendizado multimodal visa explorar as características das diversas modalidades (texto, imagem, áudio) para gerar modelos computacionais. No comércio eletrônico, devido à grande variedade das características dos produtos e à ausência ou inconsistência de informações, a combinação de informações de modos diferentes vem a ser bastante adequada. Neste trabalho são apresentados alguns experimentos para a classificação multimodal (texto e imagem) de produtos (produtos adultos) que não podem ser vendidos no marketplace da empresa parceira. Nesses experimentos, redes neurais foram usadas para treinar classificadores uni e multimodal. O classificador multimodal atingiu* 99% *de* $F1$ *contra* 98% *do modelo textual e* 94% *do visual.*

1. Introdução

O comércio eletrônico no Brasil acaba de bater recorde com um faturamento de R$$53,4$ bilhões no primeiro semestre de 2021, segundo dados da 44^{a} edição do Webshoppers, relatório elaborado pela Ebit | Nielsen[1]. Nos últimos anos, a sua importância para as empresas e a população em geral cresceu consideravelmente. Uma das razões para esse crescimento são os *marketplaces*, que durante o isolamento na situação da pandemia (entre 2020-2021) tornaram-se uma alternativa fundamental para que muitas lojas, inclusive supermercados, continuassem faturando. Segundo a Associação Brasileira de Comércio

[1]https://www.ecommercebrasil.com.br/noticias/e-commerce-no-brasil-bate-recorde-e-atinge-r-53-bilhoes-ebit-nielsen-webshoppers/

Eletrônico (ABCOMM)[2], em 2020, os *marketplaces* foram responsáveis por 78% do faturamento total do comércio eletrônico.

Em um *marketplace*, o modelo de negócio permite que um vendedor insira na plataforma de venda as informações, imagens ou descrições dos produtos personalizadas para seus anúncios. Junto com essa maior flexibilização surgem também alguns desafios como ter que lidar com a falta de informações, imagens que não satisfazem o padrão do *marketplace* e textos descritivos com especificações não estruturadas. Para tentar contornar esses problemas, algumas empresas utilizam de corretores automáticos, verificação/curadoria manual, categorizações manuais ou automáticas, ferramentas para classificação, etc.

Para realizar essa verificação/curadoria, os *marketplaces* geralmente contratam empresas especializadas ou usam plataformas de *crowdsourcing* para classificar manualmente os produtos. Contudo, devido à grande quantidade de novos produtos carregados diariamente e à natureza dinâmica das categorias, as soluções de aprendizado de máquina surgem como uma alternativa para classificar automaticamente os produtos e reduzir o custo desta tarefa [Zahavy et al. 2018].

Nesse cenário, a classificação correta de um produto é fundamental não apenas para garantir a visibilidade do produto, mas também para determinar se ele satisfaz as políticas de venda do *marketplace*. Este trabalho foca especificamente nesse último ponto, apresentando soluções para classificar automaticamente os produtos cadastrados pelos vendedores com o intuito de barrar aqueles que não podem ser vendidos devido às políticas do *marketplace* (por exemplo, produtos ilegais).

Como as informações dos produtos estão, em sua maioria, na forma de dados não estruturados nas modalidades de texto e imagem, este trabalho investigou como modelos uni e multimodal se saem na classificação de produtos da categoria Adulto.[3] A hipótese investigada neste trabalho é a de que os métodos que lidam com o aprendizado em mais de uma modalidade têm o potencial de enriquecer a representação dos produtos, tornando possível uma melhora no desempenho da tarefa de classificação em relação aos modelos unimodais. Para tanto, foram utilizados os conjuntos de dados de produtos contendo informações textuais (títulos e descrições) e visuais (fotos e imagens) fornecidos pela empresa parceira deste projeto.

As principais contribuições deste trabalho são: (i) análise de desempenho da classificação multimodal de produtos que não devem ser vendidos no *marketplace* da empresa parceira, e (ii) geração de um modelo multimodal com alto desempenho e pronto para entrar em produção.

O restante deste artigo está organizado como segue. Na Seção 2 são apresentados alguns trabalhos da literatura selecionados como os mais relevantes para tratar o problema de classificação multimodal para o *e-commerce*. A Seção 3 descreve o conjunto de dados (3.1) e os modelos unimodal textual (3.2), visual (3.3) e multimodal (3.4) desenvolvidos neste trabalho. A avaliação desses modelos é apresentada na Seção 4 e a Seção 5 encerra este documento com algumas conclusões e propostas de trabalhos futuros.

[2]`https://abcomm.org/noticias/marketplaces-crescimento-exponencial-ao-longo-da-pandemia/`

[3]A categoria Adulto contém produtos adequados para maiores de 18 anos como itens relacionados a sexo.

2. Trabalhos relacionados

O aprendizado de máquina no contexto do comércio eletrônico, assim como em diversos outros cenários reais, enfrenta desafios para lidar com uma distribuição irregular de dados. Uma das estratégias adotadas para tentar solucionar esse problema é combinar as informações vindas de diferentes modalidades, como o textual e o visual, no que chamamos de aprendizado multimodal [Bi et al. 2020]. Segundo [Peng et al. 2018], as características específicas de cada modalidade levam a uma heterogeneidade na representação, o que faz com que seja necessário refinar as modalidades para que os métodos de aprendizado multimodal não aprendam características erradas ou prejudiciais para o modelo.[4] Além disso, há a dificuldade relacionada a como aprender essas representações de maneira conjunta.

No aprendizado multimodal, as informações em um modo (por exemplo, o textual) são combinadas com as informações em outro modo (por exemplo, o visual) via um processo de fusão. Os tipos de fusão podem ser divididos com base no momento em que a fusão ocorre, podendo haver fusão no início e no fim do processamento das modalidades, e são agrupados em: fusão em nível de recurso (*early fusion*) ou fusão em nível de decisão (*late fusion*) [Zahavy et al. 2018].

Em [Bi et al. 2020], os autores utilizaram as duas estratégias e a fusão em nível de decisão obteve melhor desempenho ($F1 = 90,94\%$). Em [Chordia and Kumar 2020], os autores também fazem uso de modelos para cada modalidade específica, mas utilizam também um técnica de *co-attention* proposta em [Lu et al. 2016] , à qual atribuem uma importante contribuição para a performance geral de sua proposta ($F1 = 91,36\%$).

Os autores de [Wirojwatanakul and Wangperawong 2019] também utilizaram a abordagem *late fusion* para categorizar produtos à venda na Amazon. Eles treinaram modelos para as modalidades específicas separadamente (texto, imagem, descrição) que chamaram de modelo de fusão tri-modal. Embora tenham obtido bons resultados ($F1 = 88,2\%$) os autores apontaram um número significante de erros e um direcionamento para extensão dos dados em trabalhos futuros.

Em [Zahavy et al. 2018], os autores fizeram uma arquitetura com 3 componentes: (1) uma CNN de [Kim 2014] para o texto, (2) uma CNN de [Simonyan and Zisserman 2015] para imagem e (3) uma rede neural de decisão, que aprende a escolher qual classificação considerar entre essas duas modalidades.

O que se pode resumir da breve análise dos trabalhos relacionados apresentada nesta seção, é que, no geral, todos os modelos foram desenvolvidos para modalidades de texto e imagens, utilizando de CNNs para imagens, como VGG [Simonyan and Zisserman 2015], ResNet [He et al. 2016] ou CNN de [Kim 2014]; e alguns métodos de PLN, como BERT [Devlin et al. 2018] ou *embeddings* gerados por métodos bastante conhecidos na área, como Glove [Pennington et al. 2014], para textos.

3. Experimentos

Esta seção descreve o conjunto de dados e os modelos uni e multimodal utilizados nos experimentos.

[4]Exemplos de tarefas para o refinamento nas modalidades são: remoção de *stop-words*, números ou caracteres especiais nos textos; e ajuste de regiões de interesse ou aplicação de filtros nas imagens.

3.1. Conjunto de dados

Os dados utilizados nos experimentos apresentados neste artigo são referentes a 8.668 produtos, sendo 4.334 de conteúdo categorizado como Adulto (classe positiva, 1) e outros 4.334 produtos permitidos à venda (classe negativa, 0).

Cada instância possui: (i) uma imagem que representa o produto, (ii) seu título e (iii) sua descrição. A Figura 1 traz dois exemplos desse conjunto de dados, um da classe positiva (à esquerda) e outro da classe negativa (à direita). Esse conjunto de dados foi dividido em 60% para treinamento, 30% para validação e 10% para teste. Resultaram, então, 5.202 produtos para treinamento, 2.602 para validação e 864 de teste, todos igualmente distribuídos para as duas classes.

Título: A Peça Do Adulto Do Jogo Do Sexo De 4 Partes Ajuda A Dados Da Posição Do Sexo Para Preliminares Do Quarto Dos Pares

Descrição: Jogo de Dados Sexuais Dados Eróticos de Amor Adicionam Especiarias para Casais Solteiros Adultos. Faça um aniversário de aventureiro ou [...]

Título: Cinta Colete Corretor Postura Coluna Lombar tamanho XXL=EXTRA G

Descrição: Cinta Colete Corretor Postura Coluna Lombar tamanho XXL=EXTRA G

Figura 1. Exemplos de produtos da classe positiva e negativa

3.2. Modelo textual

Para a modelagem textual (títulos e descrições) foram geradas *word embeddings* usando o FastText [Bojanowski et al. 2017], abordagem CBOW, com 64 dimensões, sendo considerada a média dos vetores das palavras para a representação da sentença.[5] Para a geração desses *embeddings* foi utilizado um *corpus* composto por títulos e descrições de aproximadamente 7,5 milhões de produtos.

Antes de obter as *word embeddings* pelo FastText (3), todos os textos (1) são pré-processados a fim de remover *stopwords*, caracteres especiais, sequências numéricas, links e realizar a conversão para minúsculo (2), como ilustrado no exemplo abaixo.

```
1. Faca Esportiva Xingu XV2562 Outdoor com Bainha eBússola Camuflada
2. faca esportiva xingu xv outdoor bainha bussola camuflada
3. [ 0.01715468, -0.01149213, 0.03243327, ... ,-0.06800657, -0.03518752]
```

Nos experimentos para a classificação de itens da categoria Adulto foram gerados dois modelos textuais: um apenas para títulos e outro para títulos e descrições. No modelo textual de título, a entrada é o vetor de 64 dimensões das *embeddings* do título. Já no

[5]O FastText foi escolhido por ter sido o de melhor desempenho em experimentos preliminares para cálculo de similaridade multimodal com dados de comércio eletrônico, em comparação com Glove [Pennington et al. 2014] e Word2Vec [Mikolov et al. 2013]. Vale mencionar que as *word embeddings* geradas usando o FastText também foram melhores quando comparadas com as de domínio geral do NILC [Hartmann et al. 2017].

modelo de título e descrições, a entrada são dois vetores concatenados onde primeiro é do título e o segundo, das descrições, totalizando um vetor de tamanho 128 como ilustrado na Figura 2. Por não haver *overfitting* no treino dos modelos textuais, optou-se por não adicionar uma camada de regularização (dropout).

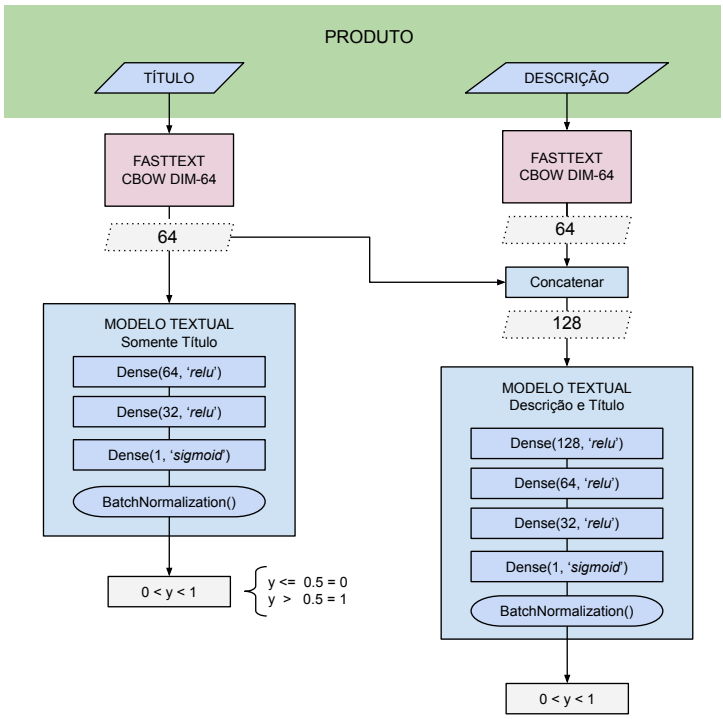

Figura 2. Descrição dos modelos unimodais textuais, onde a saída (y) representa a probabilidade do produto ser Adulto.

Para usar o modelo para a classificação pode-se considerar, por exemplo, que dado um y maior do que $0,5$ indica um produto que deve ser considerado Adulto (saída/classe igual a 1) e que abaixo de $0,5$ seja não Adulto (saída/classe igual a 0).

Todos os modelos unimodais foram implementados utilizando o Keras (v2.5.0) [Chollet et al. 2015] e seus hiper-parâmetros (como *batch-size*, *loss* e outros) foram escolhidos de maneira empírica O treinamento foi realizado com 350 épocas, tamanho do *batch* igual a 32 e uma taxa de aprendizado Adam de 1×10^{-5}, *loss binary cross-entropy*, com um total 6.277 parâmetros em cada modelo. A máquina utilizada para o treinamento dos modelos possui Windows 10 build 20H2, processador AMD Ryzen 5 3600 6-Core Processor com 12 núcleos de até 3.6GHz, 16GB de memória RAM e placa de vídeo (GPU) NVIDIA GeForce RTX2060 de 6GB de VRAM dedicado. O tempo de treinamento de cada modelo ficou em torno de 6 minutos.

3.3. Modelo visual

No modelo visual, ilustrado na Figura 3, todas as imagens foram redimensionadas para 224×224 pixeis, mantendo as 3 dimensões que representam as cores RGB. Alguns produtos possuíam mais de uma imagem, mas nos experimentos descritos neste artigo optou-se por limitar apenas à uma imagem.

Para esse modelo visual também utilizou-se o Keras [Chollet et al. 2015], que possui diversos modelos de redes neurais convolucionais pré-treinados em *datasets* como ImageNet [Deng et al. 2009], MNIST [LeCun and Cortes 2010] e CIFAR [Krizhevsky 2012]. Para a construção desse modelo, técnicas como *transfer learning* e *fine tuning* foram utilizadas para otimizar a extração de características e o aprendizado.

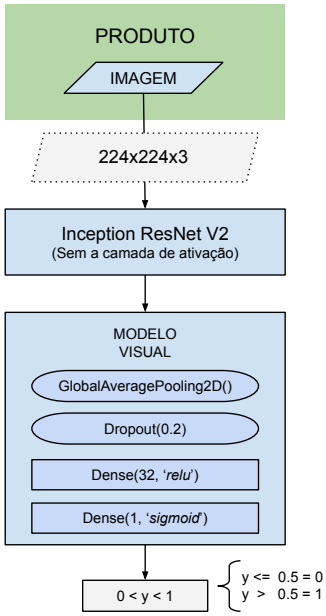

Figura 3. Descrição da combinação do modelo pré-treinado Inception ResNet V2 concatenado ao nosso modelo visual com função de ativação *sigmoid*.

Para o *transfer learning*, utilizou-se a rede Inception ResNet V2 [Szegedy et al. 2017] pré-treinada para o ImageNet de 1.000 classes. Sua arquitetura possui 780 camadas e a camada de classificação foi removida e concatenada ao modelo visual desse experimento. Para o *fine tuning*, foi feito um congelamento da camada inicial até a camada 650, a fim de garantir que o modelo faça a mesma extração de característica do seu pré-treinamento. O modelo visual (Figura 3) gerado dessa maneira é, então, aplicado para fazer a classificação com base nas *features* de uma nova imagem extraídas pela rede Inception ResNet V2.

O treinamento foi realizado com 25 épocas, tamanho do *batch* igual a 16 e uma taxa de aprendizado Adam de 1×10^{-5}, e *loss binary cross-entropy*, com um total $55.873.736$ parâmetros. A máquina usada para o treinamento desse modelo visual foi a mesma usada no modelo textual e o treinamento levou cerca de 18 minutos.

3.4. Modelo multimodal

O modelo multimodal utiliza os modelos unimodais pré-treinados sem as suas camadas de ativação. Os modelos são concatenados e, então, transportados por camadas densas. A função de ativação também é a *sigmoid* e para o treinamento foram configuradas 100 épocas, tamanho de *batch* igual a 32, *loss binary cross-entropy* e uma taxa de aprendizado Adam de 1×10^{-4}. O módulo visual é congelado até a camada 650 e o textual é congelado até a camada `Dense(32, 'relu')` (veja Figura 2).[6]

[6]Novamente, a mesma máquina foi usada para o treinamento do modelo multimodal por cerca de 1 hora e meia.

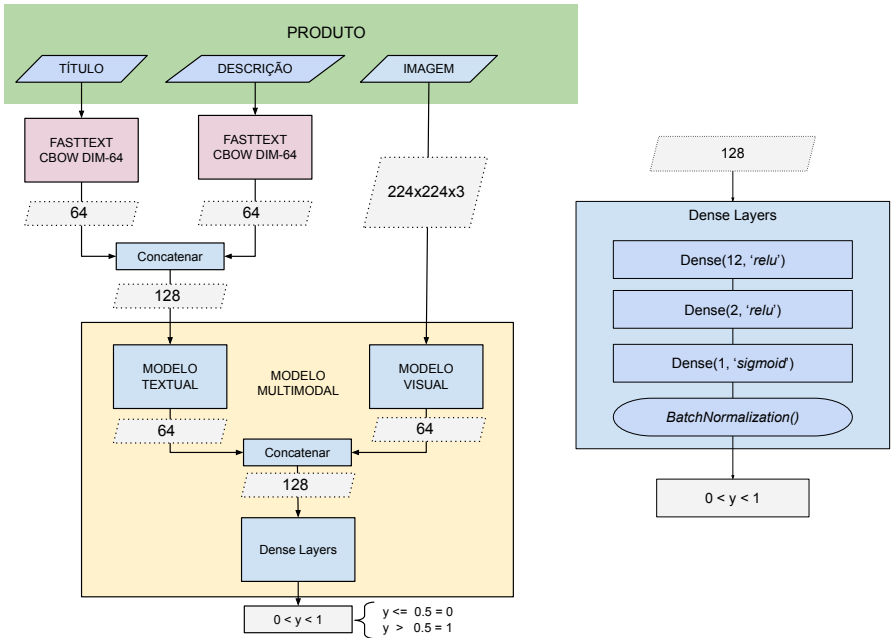

Figura 4. Descrição do modelo multimodal textual-visual produzido para este experimento.

4. Resultados

Os valores para as medidas de avaliação precisão, cobertura e $F1$, para os modelos unimodais e multimodal, para cada classe, são apresentados na Tabela 1.

Tabela 1. Valores das medidas de avaliação para os modelos unimodais de texto e de imagem e multimodal

modelo	classe	precisão	cobertura	$F1$
`textual-título`	0	93%	97%	95%
`textual-título`	1	96%	93%	94%
`textual-título+descrição`	0	98%	98%	98%
`textual-título+descrição`	1	98%	98%	98%
`visual`	0	94%	94%	94%
`visual`	1	94%	94%	94%
`multimodal`	0	99%	99%	99%
`multimodal`	1	99%	99%	99%

O modelo visual foi o que apresentou pior desempenho quando comparado aos demais. Acredita-se que os modelos textuais podem dar foco em palavras específicas e, com isso, podem ocasionar a classificação correta de um produto adulto. Já no visual, a variação de imagens de diferentes tipos de produtos pode ter trazido ruído para o aprendizado. Contudo, vale ressaltar que os modelos unimodais, quando fundidos no modelo multimodal, levaram a um aumento nos valores das medidas, indicando que os mesmos são complementares e comprovando nossa hipótese.

Para facilitar a análise qualitativa dos resultados, uma interface de visualização foi implementada. Nessa interface, ilustrada na Figura 5, são exibidos os erros associados aos modelos que erraram (indicado pelas cores). Por exemplo, o primeiro produto foi classificado de modo errado como positivo (produto adulto) por todos os modelos, enquanto o

terceiro exemplo também foi equivocadamente classificado como positivo pelos modelos visual e multimodal. Cada exemplo é acompanhado da classe correta (entre colchetes), da classe atribuída de modo errado pelo classificador, e as informações do produto usadas na classificação: título, descrição e imagem.

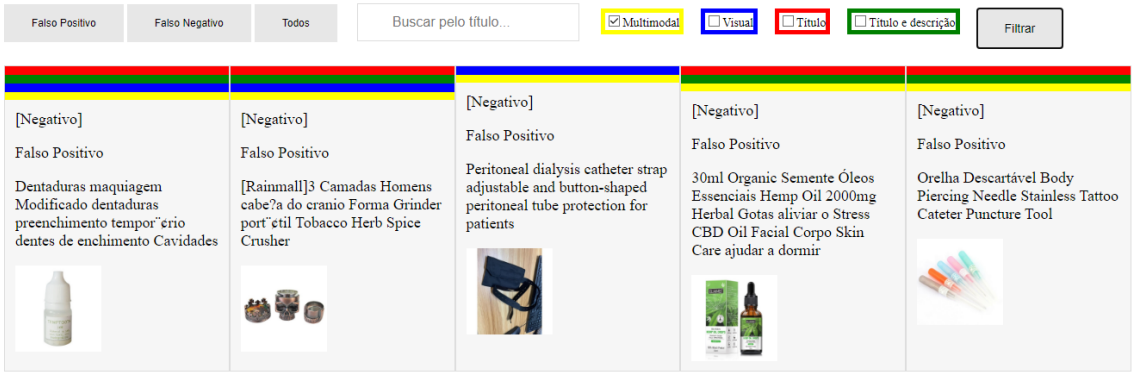

Figura 5. Interface desenvolvida para análise qualitativa dos erros de classificação dos modelos treinados ilustrando alguns falsos positivos.

Apenas 4 produtos que não podem ser vendidos no *marketplace* da empresa parceira foram classificados de forma errada (falso negativo) por todos os modelos.[7] Após analisar os possíveis motivos desse erro, notou-se que algumas informações textuais desses produtos estavam em inglês. Essa diferença de idioma pode ter sido prejudicial porque as *word embeddings* não foram treinadas para produtos nessa língua. Também foi possível observar que geralmente os produtos adultos possuem palavras específicas que podem ser utilizadas para classificar o produto nessa categoria. Dessa observação surge a proposta de retreinar o modelo do FastText incluindo produtos proibidos para venda, pois as *word embeddings* foram treinadas apenas a partir de produtos que estão disponíveis para a venda.

5. Conclusões e Trabalhos futuros

Este trabalho avaliou a classificação multimodal de produtos que não devem ser vendidos no *marketplace* da empresa parceira. Nos experimentos, constatou-se que o modelo unimodal de título e descrição apresentou um resultado muito bom ($F1 = 98\%$) mas sua combinação com o modelo visual, no modelo multimodal, foi ainda melhor ($F1 = 99\%$).

Como propostas de trabalhos futuros, tem-se: (i) retreinar o modelo do FastText incluindo itens do conjunto de produtos proibidos para venda; (ii) investigar a abordagem *ensemble* dos modelos e outras opções de fusão; (iii) estender os experimentos para outras categorias de produtos proibidos e (iv) colocar o modelo gerado em produção no *marketplace* da empresa parceira.

Agradecimentos

Esse artigo e a pesquisa desenvolvida não seriam possíveis sem o apoio da americanas s.a. Digital Lab, especialmente o apoio de José Pizani, Ester Campos e Jonas Ferreira. Esse trabalho é parte do projeto “Dos dados ao conhecimento: extração e representação de informação no domínio do e-commerce” (Projeto de extensão - UFSCar #23112.000186/2020-97).

[7]Essas instâncias não são apresentadas neste artigo por considerarmos seu conteúdo impróprio para o público em geral.

Referências

Bi, Y., Wang, S., and Fan, Z. (2020). A multimodal late fusion model for e-commerce product classification. *Proceedings of The 2020 SIGIR Workshop On eCommerce*, abs/2008.06179.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Chordia, V. and Kumar, V. (2020). Large scale multimodal classification using an ensemble of transformer models and co-attention. *CoRR*, abs/2011.11735.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Krizhevsky, A. (2012). Learning multiple layers of features from tiny images. *University of Toronto*.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289–297, Red Hook, NY, USA. Curran Associates Inc.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Peng, Y., Qi, J., and Yuan, Y. (2018). Modality-specific cross-modal similarity measurement with recurrent attention network. *Trans. Img. Proc.*, 27(11):5585–5599.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4278–4284. AAAI Press.

Wirojwatanakul, P. and Wangperawong, A. (2019). Multi-label product categorization using multi-modal fusion models. *CoRR*, abs/1907.00420.

Zahavy, T., Krishnan, A., Magnani, A., and Mannor, S. (2018). Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).