# A multimodal deep learning architecture for smoking detection with a small data approach

Róbert Lakatos[2,3,4], Péter Pollner[1], András Hajdu[2], and Tamás Joó[1,4]

[1]Data-Driven Health Division of National Laboratory for Health Security, Health Services Management Training Centre, Semmelweis University

[2]Department of Data Science and Visualization, Faculty of Informatics, University of Debrecen

[3]Doctoral School of Informatics, University of Debrecen

[4]Neumann Technology Platform, Neumann Nonprofit Ltd.

## Abstract

**Introduction:** Covert tobacco advertisements often raise regulatory measures. This paper presents that artificial intelligence, particularly deep learning, has great potential for detecting hidden advertising and allows unbiased, reproducible, and fair quantification of tobacco-related media content.

**Methods:** We propose an integrated text and image processing model based on deep learning, generative methods, and human reinforcement, which can detect smoking cases in both textual and visual formats, even with little available training data.

**Results:** Our model can achieve 74% accuracy for images and 98% for text. Furthermore, our system integrates the possibility of expert intervention in the form of human reinforcement.

**Conclusions:** Using the pre-trained multimodal, image, and text processing models available through deep learning makes it possible to detect smoking in different media even with few training data.

**Keywords:** AI-supported preventive healthcare, pre-training with generative AI, multimodal deep learning, automated assessment of covert advertisement, few-shot learning; smoking detections

# 1 Introduction

The WHO currently estimates that smoking causes around 8 million deaths a day. It is the leading cause of death from a wide range of diseases, for example, heart attacks, obstructive pulmonary disease, respiratory diseases, and cancers. 15% of people aged 15 years and over smoke in the OECD countries and 17% in the European Union.[1] Moreover, of the 8 million daily deaths, 15% result from passive smoking.[2] The studies[3,4] below highlight the influence of smoking portrayal in movies and the effectiveness of health communication models. However, quantifying media influence is complex. For internet media like social sites, precise ad statistics are unavailable. Furthermore, calculating incited and unmarked ads poses a significant difficulty as well. Therefore, accurate knowledge of the smoking-related content appearing in individual services can be an effective tool in reducing the popularity of smoking. Methods for identifying content include continuous monitoring of advertising intensity,[5] structured data generated by questionnaires,[6] and AI-based solutions that can effectively support these goals. The authors of the article "Machine learning applications in tobacco research"[7] point out in their review that artificial intelligence is a powerful tool that can advance tobacco control research and policy-making. Therefore, researchers are encouraged to explore further possibilities.

Nonetheless, these methods are highly data-intensive. In the case of image processing, an excellent example of this is the popular ResNet[8] image processing network, which was trained on the ImageNet dataset[9] containing 14,197,122 images. Regarding text processing, we can mention the popular and pioneering BERT network[10] trained by the Toronto BookCorpus[11] was trained by the 4.5 GB of Toronto BookCorpus. Generative text processing models such as GPT[12] are even larger and were trained with significantly more data than BERT. For instance, the training set of GPT 3.0 was the CommonCrawl[13] dataset, which has a size of 570 GB.

The effective tools for identifying the content of natural language texts are topic modeling[14] and the embedding of words,[15–17] tokens, sentences,[18] or characters[19] clustering.[20] For a more precise identification of the content elements of the texts, we can use the named-entity recognition[21] techniques. In image processing, we can highlight classification and object detection to detect smoking. The most popular image processing models are VGG,[22] ResNet,[8] Xception,[23] EfficientNet,[24] Inception,[25] and YOLO.[26] Moreover, there are architectures like CAMFFNet,[27] which are specifically recommended for smoking detection. The development of multimodal models also is gaining increasing focus,[28,29] which can use texts and images the solve the

tasks at the same time. For movies, scene recognition is particularly challenging compared to images.[30] Scene recognition is also linked to sensitive events such as fire, smoke, or other disaster detection systems,[31] but there are attempts to investigate point-of-sale and tobacco marketing practices[32] as well.

We concluded that there is currently no publicly available specific smoking-related dataset that would be sufficient to train a complex model from scratch. Hence, we propose a multimodal architecture that uses pre-trained image and language models to detect smoking-related content in text and images. By combining image processing networks with multimodal architectures and language models, we leverage textual and image data simultaneously. This offers a data-efficient and robust solution that can be further improved with expert input. This paper demonstrates the remarkable potential of artificial intelligence, especially deep learning, for the detection of covert advertising, alongside its capacity to provide unbiased, replicable, and equitable quantification of tobacco-related media content.

# 2 Methods

## 2.1 Model Architecture

As illustrated in Figure 1 by a schematic flow diagram, our solution relies on pre-trained language and image processing models and can handle both textual and image data.

The first step of our pipeline is to define the incoming data format because need to direct the data to the appropriate model for its format. The video recordings are analyzed with multimodal and image processing models, while the texts are analyzed with a large language model. In the case of video recordings, we applied the CLIP-ViT-B-32 multilingual[33,34] model. The model has been developed for over 50 languages with a special training technique.[33] The model supports Hungarian, which was our target language. We use the CLIP-ViT-B-32 model as a filter. After filtering, to achieve more accurate results, we recommend using the pre-trained EfficientNet B5 model, which we fine-tuned with smoking images for the classification task.

To process texts, we use name entity recognition to identify smoking-related terms. For this purpose, we have integrated into our architecture an XLM-RoBERTa model[35] that is pre-trained, multilingual, and also supports the Hungarian language, which is important to us.

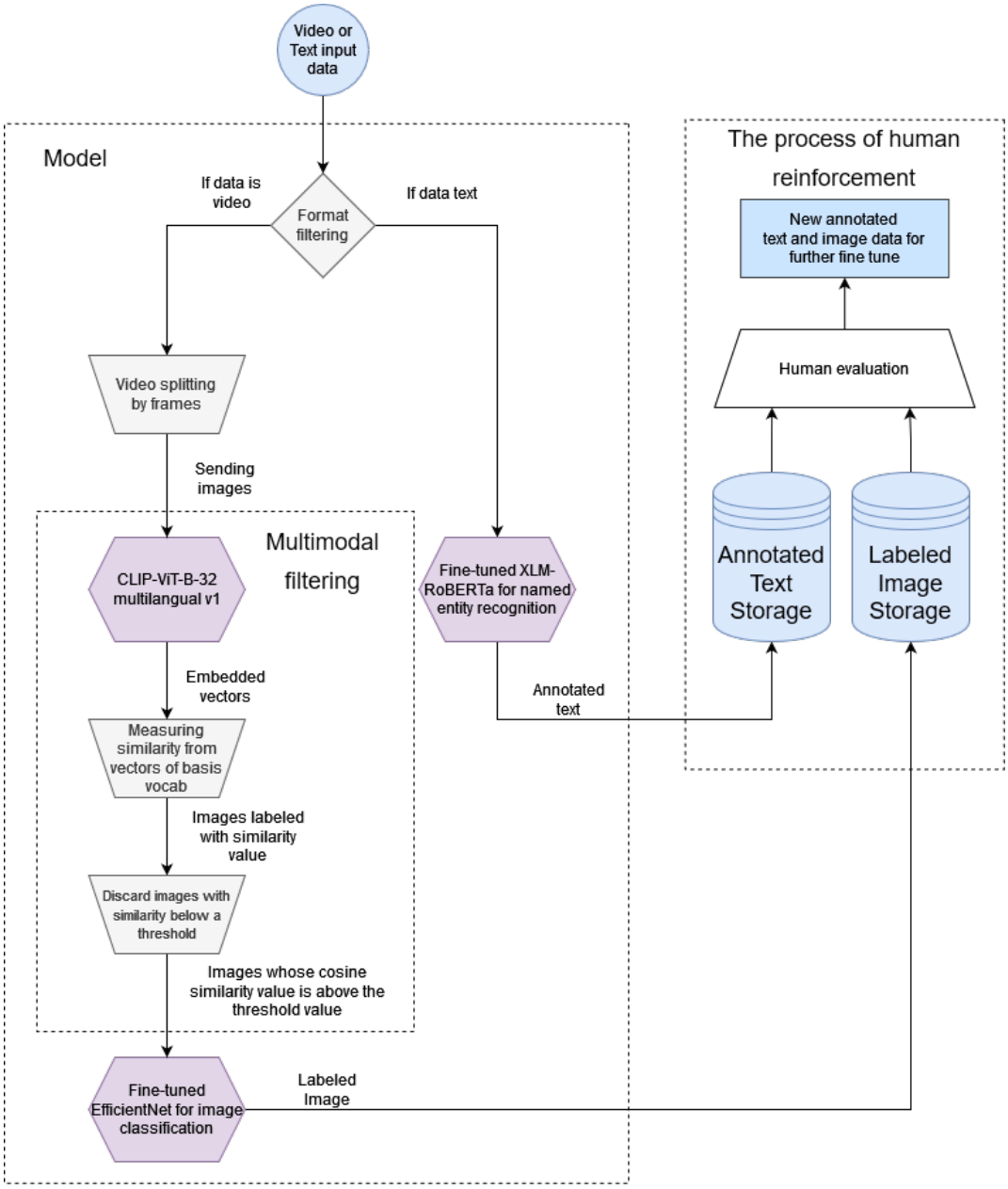


Figure 1: Schematic flow diagram of the architecture.

## 2.2 Format check

The first step in processing is deciding whether the model has to process video recordings or text data. Since there are many formats for videos and texts, we chose the simple solution of only supporting mp4 and txt file formats. The mp4 is a popular video format, and practically all other video recording formats can be converted to mp4. We consider txt files utf8-encoded raw text files that are ideally free of various metadata. It is important to emphasize that here we ignore the text cleaning processes required to prepare raw text files. The reason is that we did not deal with faulty or txt files requiring further cleaning during the trial.

## 2.3 Processing of videos and images

The next step in the processing of processing video footage is to break it down into frames by sampling every second. The ViT image encoder of the CLIP-ViT-B-32 model was trained by its creators for various image sizes. For this, they used the ImageNet[9] dataset in which the images have an average size of 469×387 pixels.

The developers of CLIP-ViT-B-32 do not recommend an exact resolution for the image encoder. The model specification only specifies a minimum resolution of 224×224. In the case of EfficientNetB5, the developers have optimized an image size of 224×224. For these reasons, we have taken this image size as a reference and transformed the images sampled from the video recordings to this image size.

## 2.4 Multimodal filtering

The images sampled from the video recordings were filtered using the CLIP-ViT-B-32 multilingual v1 model. The pre-trained CLIP-ViT-B-32 multilingual v1 model consists of two main components from a ViT[36] image processing model and a DistilBERT-based[37] multilingual language model. We convert into a 512-long embedded vector[16] the images and texts with CLIP-ViT-B-32. The embedded vectors for texts and images can be compared based on their content meaning if we measure cosine similarities between the vectors. The cosine similarity is a value falling in the interval [-1,1], and the similarity of two vectors will be larger the closer their cosine similarity is to 1.

Since we aimed to find smoking-related images, we defined a smoking-related term. We converted it to a vector and measured it against the embedded vectors generated from the video images. The term we chose was the word ”smoking”. We can use more complex expressions, which could complicate the measurement results interpretation.

The cosine similarity of the vectors produced by embedding the images always results in a scalar value compared to the vector created from our expression related to ”smoking”. However, the decision limit between the distances measured between the vectors produced by the CLIP-ViT-B-32 model is not always clear. Namely, even in the case of images with meanings other than ”smoking”, we get a value that is not too distant.

We had to understand the distribution of the smoking images to eliminate this kind of blurring

of the decision boundary. To this end, we examined the characteristics of the distribution of the images. It is clear from Figure 2 that because the images with a semantic meaning closer to smoking appear randomly in a video recording, it is difficult to grasp the series of images that can be useful for us. Figure 2 is actually a function whose vertical axis has the cosine similarity values belonging to the individual images. At the same time, the horizontal axis shows the position of the images in the video. To solve this problem, we introduced the following procedure. If we put the cosine similarity values in ascending order, we get a function that describes the ordered evolution of the cosine similarity values.

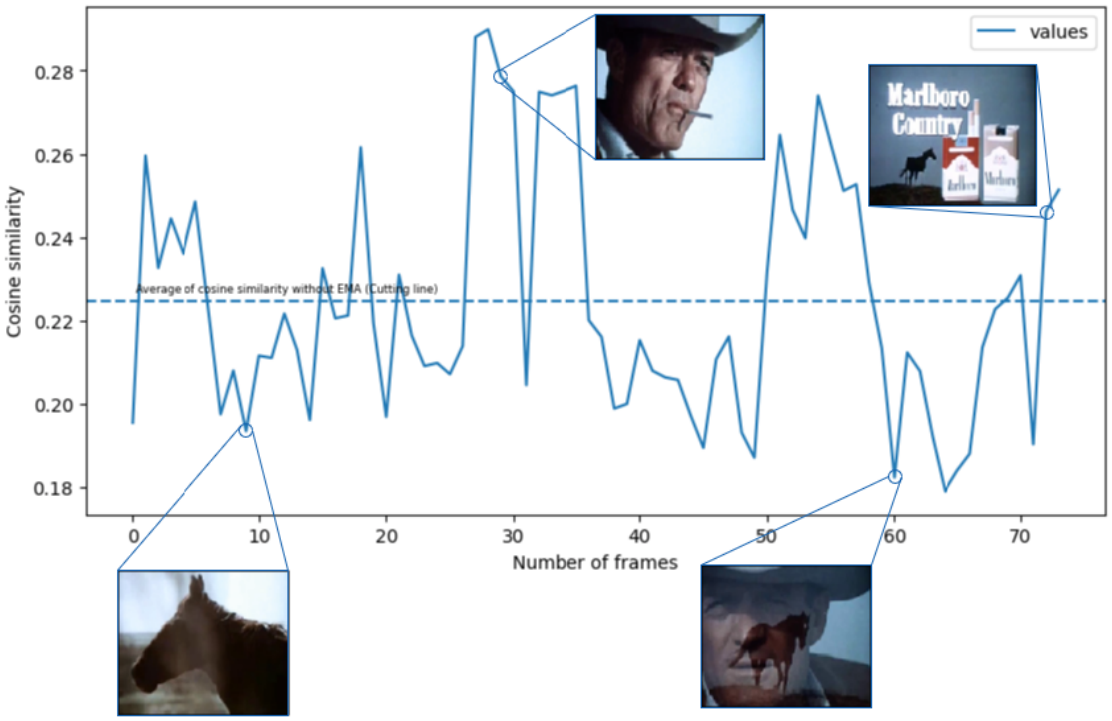


Figure 2: The cosine similarity of the images obtained from the video recording in chronological order.

The ordered function generated from Figure 2 can be seen in Figure 3. As shown in Figures 2 and 3, we found that if we take the similarity value of the images sampled from the given sample to the word ”smoking”, their average results in a cutting line, and we can use it as a filter.

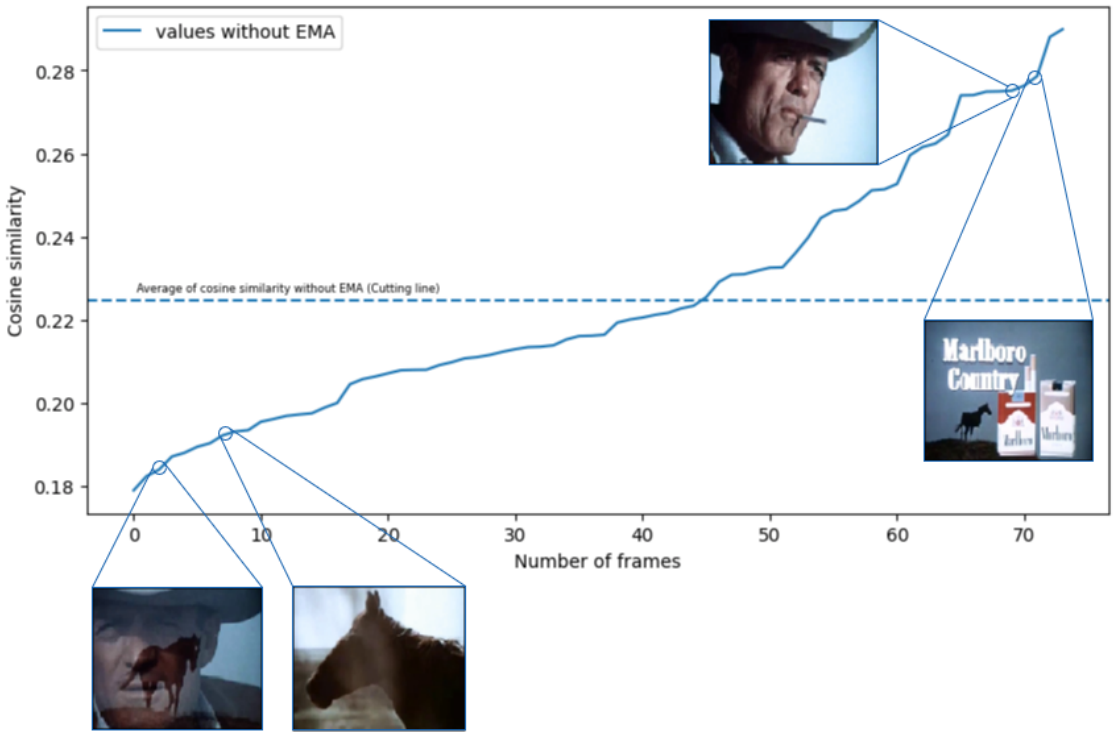


Figure 3: The images are in an orderly manner based on the cosine similarity values.

Furthermore, considering the specifics of the video recordings, we consider that the average can be corrected with a constant value. In this mean, the constant value can thus also be defined as the hyperparameter of the model. We chose the 0 default value for the correction constant because of more apparent measurements. Because the choice of the best constant value may differ depending on the recording type and may distort the exact measurement results.

## 2.5 Fine-tuned image classification

After filtering the image set with a multimodal model, we applied an image processing model to classify the remaining images further to improve accuracy. Among the publicly available datasets on smoking, we have used the "smoker and non-smoker"[38] for augmented[39] fine-tuning. We selected the following models for the task. EfficientNet, Inception, ResNet, VGG, and Xception. The EfficientNet B5 version was the best, with an accuracy of 93.75%. Table S1 of the supplemental contains our detailed measurement results concerning all models.

## 2.6 Processing of text

In the case of detecting smoking terms in texts, we approached the problem as an NER task and focused on the Hungarian language. Since we could not find a dataset containing annotated smoking phrases available in Hungarian. Therefore, to generate the annotated data, we used the generational capabilities of ChatGPT, the smoking-related words of the Hungarian synonyms and antonyms dictionary,[40] and prompt engineering. Accordingly, we selected words related to smoking from the synonyms and antonyms dictionary and asked ChatGPT to suggest further smoking-related terms besides words from the Hungarian dictionary. Finally, we combined the synonyms and the expressions generated by ChatGPT into a single dictionary.

We created blocks of a maximum of 5 elements from the words in our dictionary. Each block contained a random combination of a maximum of 5 words. The blocks are disjoint, so they do not contain the same words. This mixing step was done 10 times. This means that, in one iteration, we could form 8 blocks of 5-element disjunct random blocks from our 43-word dictionary. By doing all these 10 times, we produced 80 blocks. However, due to the 10 repetitions, the 80 blocks were no longer disjoint. In other words, if we string all the blocks together, we get a dictionary in which every synonym for smoking appears a maximum of 10 times.

We made a prompt template to which, by attaching each block, we instructed ChatGPT to generate texts containing the specified expressions. Since ChatGPT uses the Hungarian language well, the generated texts contained our selected words by the rules of the Hungarian language, with the correct conjugation. An example of our prompts is illustrated in Table 1.

Table 1: A 3 elements example prompt for ChatGPT.

| |
|---|
| Generate a short text about smoking. |
| The text strictly contains the following words in the different sentences: |
| smoking, tobacco, cigar |

We did not specify how long texts should be generated by ChatGPT or that every word of a 5-element block should be included in the generated text. When we experimented with ChatGPT generating fixed-length texts, it failed. Therefore, we have removed the requirement for this.

Using this method, we created a smoking-related corpus consisting of 80 paragraphs, 49000 characters, and 7160 words. An English example of a generated text is presented in Table 2.

Table 2: An example paragraph generated by from the prompt of Table 1.

| |
|---|
| Smoking is a widespread and addictive habit that involves inhaling and exhaling the smoke produced by burning tobacco. Whether it's a hand-rolled cigar or a manufactured cigarette, the act of smoking revolves around the consumption of tobacco. Despite the well-known health risks, many individuals continue to engage in smoking due to its addictive nature. The allure of a cigar or a cigarette can be strong, making it challenging for people to quit smoking even when they are aware of its detrimental effects. Education and support are crucial in helping individuals break free from the cycle of smoking and its associated harms. |

To find the best model according to the possibilities of our computing environment and the support of the Hungarian language, we tested the following models: XLM RoBERTa base and large, DistilBERT base cased, huBERT base,[41] BERT base multilingual,[42] Sentence-BERT.[43] The best model was the XLM RoBERTa large one, which achieved 98% accuracy and 96% F1-score on the validation dataset and an F1-score of 91% with an accuracy of 98% on the test dataset.

## 2.7 Human reinforcement

In the architecture we have outlined, the last step in dealing with the lack of data is to ensure the system's continuous development capability. For this, we have integrated human confirmation into our pipeline. The essence is that our system's hyperparameters should be adjustable and optimizable during operation and that the data generated during detection can be fed back for further fine-tuning. The cutting line used in multimodal filtering is a hyperparameter of our model. As a result, a more accurate result can be achieved by using human confirmation during the operation. The tagged images and annotated texts from the processed video recordings and texts are transferred to permanent storage in the last step of the process. This dynamically growing dataset can be further validated with additional human support, and possible errors can be filtered. So, False positives and False negatives can be fed back into the training datasets.

# 3 Results

We collected video materials to test the image processing part of our architecture. The source of the video materials was the video-sharing site YouTube. Taking into account the legal rules regarding the usability of YouTube videos, we have collected 5 pieces short advertising films from the Malboro and Philip Moris companies. We ensured not to download videos longer than 2 minutes because longer videos, such as movies, would have required a special approach and additional pre-processing. Furthermore, we downloaded the videos at 240p resolution and divided them into frames by sampling every second. Each frame was transformed to a resolution of 224×224 pixels. We manually annotated all videos. The downloaded videos averaged 64 seconds and contained an average of 13 seconds of smoking.

With the multimodal filtering technique, we discarded the images that did not contain smoking. Multimodal filtering found 25 seconds of smoking on average in the recording. The accuracy of the identified images was 62%. The multimodal filtering could filter out more than half of the 64-second, on average, videos. We also measured the performance of the fine-tuned EfficientNet B5 model by itself. The model detected an average of 28 seconds of smoking with 60% accuracy. We found that the predictions of the two constructions were sufficiently diverse to connect them using the boosting ensemble[44] solution. By connecting the two models, the average duration of perceived smoking became 12 seconds with 4 seconds on average error and 74% accuracy. The ensemble solution was the best approach since the original videos contained an average of 13 seconds of smoking. We deleted the videos after the measurements and did not use them anywhere for any other purpose.

We created training and validation datasets from Hungarian synonyms for smoking using ChatGPT. We trained our chosen large language models until their accuracy on the validation dataset did not increase for at least 10 epochs. The XLM-RoBERTa model achieved the best performance on the validation dataset with an F1-score of 96% and 98% accuracy. For the final measurement, we created test data from an online text related to smoking by manual annotation.[45] The text of the entire test data is included in the Table S20 supplemental. The fine-tuned XLM-RoBERTa model achieved 98% accuracy and 0.91 F1 score on the test dataset.

# 4 Conclusions

Multimodal and image classification models are powerful for classification tasks. In return, however, they are complex and require substantial training data, which can reduce their explainability and usability. In turn, our solution showed that pre-trained multimodal and image classification models exist that allow smoking detection even with limited data and in the matter of low-resource languages if we use the potential of human reinforcement, generative, and ensemble methods. In addition, we see further development opportunities if our approach is supplemented with an object detector, which can determine the time of occurrence of objects and their position. Moreover, with the expected optimization of the automatic generation of images in the future and the growth of the available computing power, our method used for texts can work in the case of images.

# Funding

The project no. KDP-2021 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development, and Innovation Fund, financed under the C1774095 funding scheme. Also, this work was partly funded by the project GINOP-2.3.2-15-2016-00005 supported by the European Union, co-financed by the European Social Fund, and by the project TKP2021-NKTA-34, implemented with the support provided by the National Research, Development, and Innovation Fund of Hungary under the TKP2021-NKTA funding scheme. In addition, the study received further funding from the National Research, Development and Innovation Office of Hungary grant (RRF-2.3.1-21-2022-00006, Data-Driven Health Division of National Laboratory for Health Security).

# References

[1] for Economic Co-operation, O.; Development Daily smokers (indicator). 2023.

[2] Organization, W. H. Tobacco. 2022.

[3] Chapman, S.; Davis, R. M. *Tobacco Control* **1997**, *6*, 269–271.

[4] Pechmann, C.; Shih, C. *Irvine, California: Graduate School of Management, University of California, Irvine* **1996**,

[5] Kong, G.; Schott, A. S.; Lee, J.; Dashtian, H.; Murthy, D. *Tobacco Control* **2022**,

[6] Fielding, R.; Chee, Y.; Choi, K.; Chu, T.; Kato, K.; Lam, S.; Sin, K.; Tang, K.; Wong, H.; Wong, K. *Journal of Public Health* **2004**, *26*, 24–30.

[7] Fu, R.; Kundu, A.; Mitsakakis, N.; Elton-Marshall, T.; Wang, W.; Hill, S.; Bondy, S. J.; Hamilton, H.; Selby, P.; Schwartz, R.; others *Tobacco Control* **2023**, *32*, 99–109.

[8] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

[9] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. 2009; pp 248–255.

[10] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *arXiv preprint arXiv:1810.04805* **2018**,

[11] Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision. 2015; pp 19–27.

[12] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; others Improving language understanding by generative pre-training. 2018.

[13] Common Crawl. 2022; Accessed: 2022-06-01.

[14] Blei, D. M.; Ng, A. Y.; Jordan, M. I. *Journal of machine Learning research* **2003**, *3*, 993–1022.

[15] Pennington, J.; Socher, R.; Manning, C. D. GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP). 2014; pp 1532–1543.

[16] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. *Advances in neural information processing systems* **2013**, *26*.

[17] Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. *Transactions of the association for computational linguistics* **2017**, *5*, 135–146.

[18] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:1908.10084* **2019**,

[19] Clark, J. H.; Garrette, D.; Turc, I.; Wieting, J. *Transactions of the Association for Computational Linguistics* **2022**, *10*, 73–91.

[20] Arthur, D.; Vassilvitskii, S. *k-means++: The advantages of careful seeding*; 2006.

[21] Ali, S.; Masood, K.; Riaz, A.; Saud, A. Named Entity Recognition using Deep Learning: A Review. 2022 International Conference on Business Analytics for Technology and Security (ICBATS). 2022; pp 1–7.

[22] Simonyan, K.; Zisserman, A. *arXiv preprint arXiv:1409.1556* **2014**,

[23] Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; pp 1251–1258.

[24] Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning. 2019; pp 6105–6114.

[25] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 2818–2826.

[26] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 779–788.

[27] Lin, J.; Chen, Y.; Pan, R.; Cao, T.; Cai, J.; Yu, D.; Chi, X.; Cernava, T.; Zhang, X.; Chen, X. *Computers and Electronics in Agriculture* **2022**, *202*, 107390.

[28] Liu, Y.; Guo, Y.; Liu, L.; Bakker, E. M.; Lew, M. S. *Pattern Recognition* **2019**, *93*, 365–379.

[29] Liu, Z.; Chen, F.; Xu, J.; Pei, W.; Lu, G. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**,

[30] Rao, A.; Xu, L.; Xiong, Y.; Xu, G.; Huang, Q.; Zhou, B.; Lin, D. A local-to-global approach to multi-modal movie scene segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020; pp 10146–10155.

[31] Gagliardi, A.; de Gioia, F.; Saponara, S. *Journal of Real-Time Image Processing* **2021**, *18*, 2085–2095.

[32] Bianco, F.; Moffett, C.; Abunku, P.; Chaturvedi, I.; Chen, G.; Dobler, G.; Sobolevsky, S.; Kirchner, T.; others *Authorea Preprints* **2021**,

[33] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:2004.09813* **2020**,

[34] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others Learning transferable visual models from natural language supervision. International conference on machine learning. 2021; pp 8748–8763.

[35] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. *CoRR* **2019**, *abs/1911.02116*.

[36] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others *arXiv preprint arXiv:2010.11929* **2020**,

[37] Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. *ArXiv* **2019**, *abs/1910.01108*.

[38] Khan, A. *Mendeley Data* **2020**, *1*.

[39] Shorten, C.; Khoshgoftaar, T. M. *Journal of big data* **2019**, *6*, 1–48.

[40] Viola, T. *Ellentétes jelentésű szavak adatbázisa*; Tinta Könyvkiadó, 2012.

[41] Nemeskey, D. M. Introducing `huBERT`. XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021). Szeged, 2021; p TBA.

[42] Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. *CoRR* **2018**, *abs/1810.04805*.

[43] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:1908.10084* **2019**,

[44] Dietterich, T. G. Ensemble methods in machine learning. Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1. 2000; pp 1–15.

[45] Center, H. P. Egészség Elvitelre. 2023.