

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

Róbert Lakatos^{2,3,4}, Péter Pollner¹, András Hajdu², and Tamás Joó^{1,4}

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

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

³Doctoral School of Informatics, University of Debrecen

⁴Neumann 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

27 **1 Introduction**

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

43 Nonetheless, these methods are highly data-intensive. In the case of image processing, an
44 excellent example of this is the popular ResNet⁸ image processing network, which was trained
45 on the ImageNet dataset⁹ containing 14,197,122 images. Regarding text processing, we can
46 mention the popular and pioneering BERT network¹⁰ trained by the Toronto BookCorpus¹¹ was
47 trained by the 4.5 GB of Toronto BookCorpus. Generative text processing models such as GPT¹²
48 are even larger and were trained with significantly more data than BERT. For instance, the train-
49 ing set of GPT 3.0 was the CommonCrawl¹³ dataset, which has a size of 570 GB.

50 The effective tools for identifying the content of natural language texts are topic modeling¹⁴
51 and the embedding of words,¹⁵⁻¹⁷ tokens, sentences,¹⁸ or characters¹⁹ clustering.²⁰ For a more
52 precise identification of the content elements of the texts, we can use the named-entity recogni-
53 tion²¹ techniques. In image processing, we can highlight classification and object detection to
54 detect smoking. The most popular image processing models are VGG,²² ResNet,⁸ Xception,²³
55 EfficientNet,²⁴ Inception,²⁵ and YOLO.²⁶ Moreover, there are architectures like CAMFFNet,²⁷
56 which are specifically recommended for smoking detection. The development of multimodal
57 models also is gaining increasing focus,^{28,29} which can use texts and images the solve the

58 tasks at the same time. For movies, scene recognition is particularly challenging compared
59 to images.³⁰ Scene recognition is also linked to sensitive events such as fire, smoke, or other
60 disaster detection systems,³¹ but there are attempts to investigate point-of-sale and tobacco
61 marketing practices³² as well.

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

71 **2 Methods**

72 **2.1 Model Architecture**

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

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

83 To process texts, we use name entity recognition to identify smoking-related terms. For this
84 purpose, we have integrated into our architecture an XLM-RoBERTa model³⁵ that is pre-trained,
85 multilingual, and also supports the Hungarian language, which is important to us.

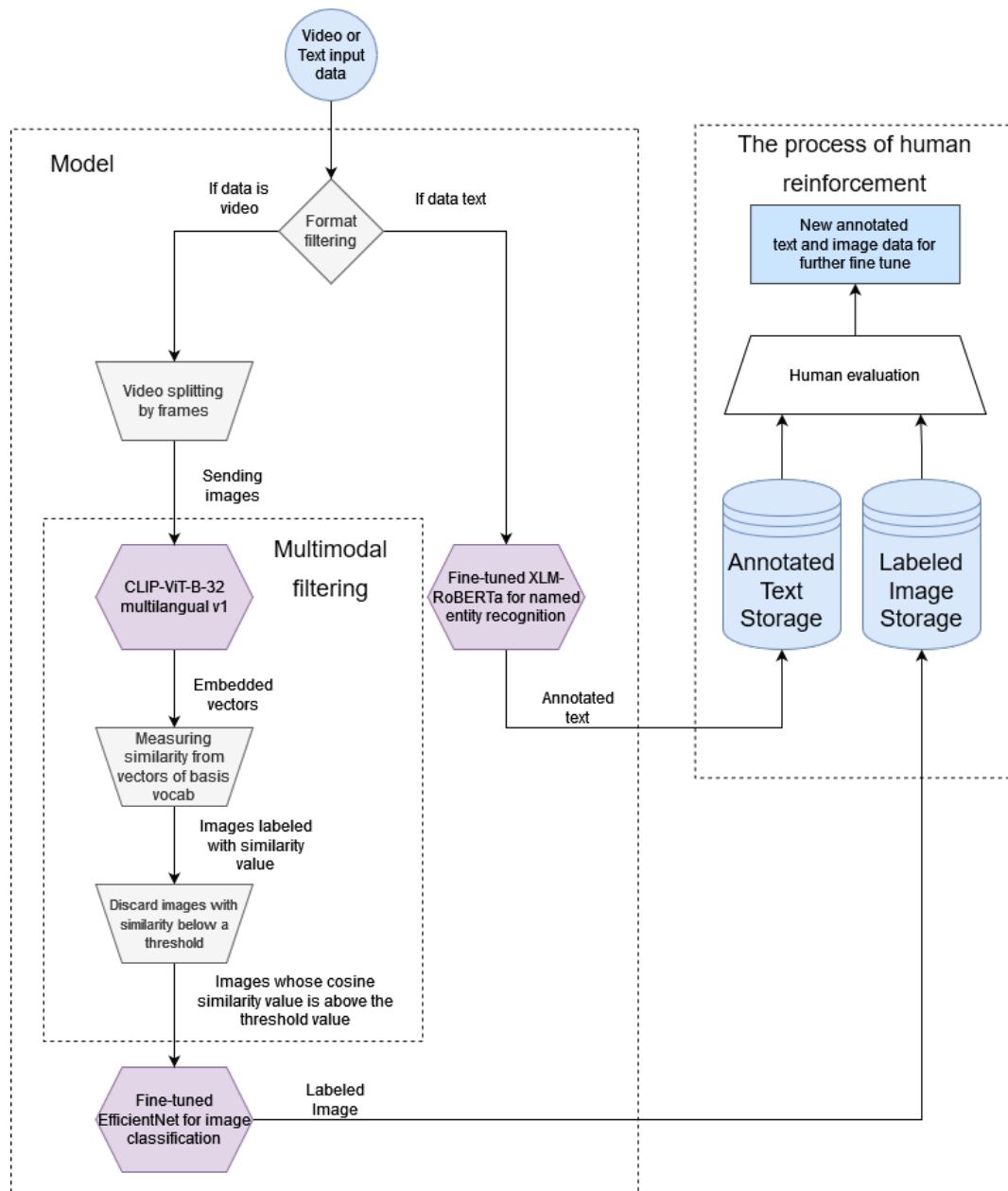


Figure 1: Schematic flow diagram of the architecture.

86 2.2 Format check

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

94 **2.3 Processing of videos and images**

95 The next step in the processing of processing video footage is to break it down into frames by
96 sampling every second. The ViT image encoder of the CLIP-ViT-B-32 model was trained by its
97 creators for various image sizes. For this, they used the ImageNet⁹ dataset in which the images
98 have an average size of 469×387 pixels.

99 The developers of CLIP-ViT-B-32 do not recommend an exact resolution for the image encoder.
100 The model specification only specifies a minimum resolution of 224×224 . In the case of Effi-
101 cientNetB5, the developers have optimized an image size of 224×224 . For these reasons, we
102 have taken this image size as a reference and transformed the images sampled from the video
103 recordings to this image size.

104 **2.4 Multimodal filtering**

105 The images sampled from the video recordings were filtered using the CLIP-ViT-B-32 multilin-
106 gual v1 model. The pre-trained CLIP-ViT-B-32 multilingual v1 model consists of two main com-
107 ponents from a ViT³⁶ image processing model and a DistilBERT-based³⁷ multilingual language
108 model. We convert into a 512-long embedded vector¹⁶ the images and texts with CLIP-ViT-
109 B-32. The embedded vectors for texts and images can be compared based on their content
110 meaning if we measure cosine similarities between the vectors. The cosine similarity is a value
111 falling in the interval $[-1,1]$, and the similarity of two vectors will be larger the closer their cosine
112 similarity is to 1.

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

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

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

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

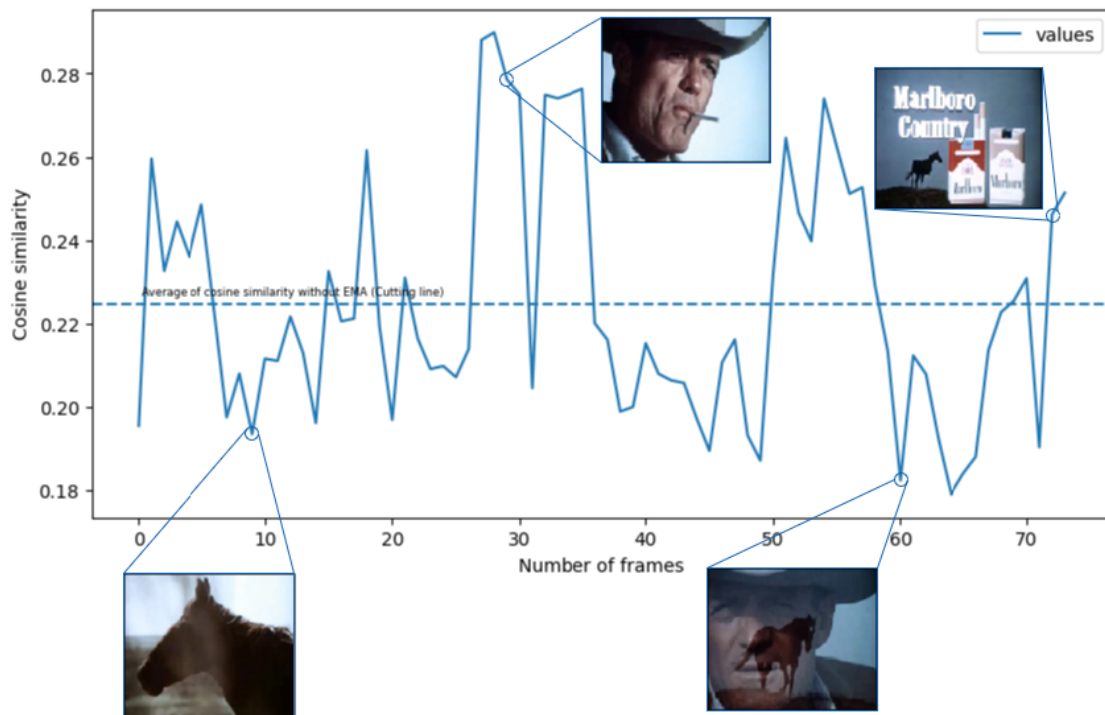


Figure 2: The cosine similarity of the images obtained from the video recording in chronological order.

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

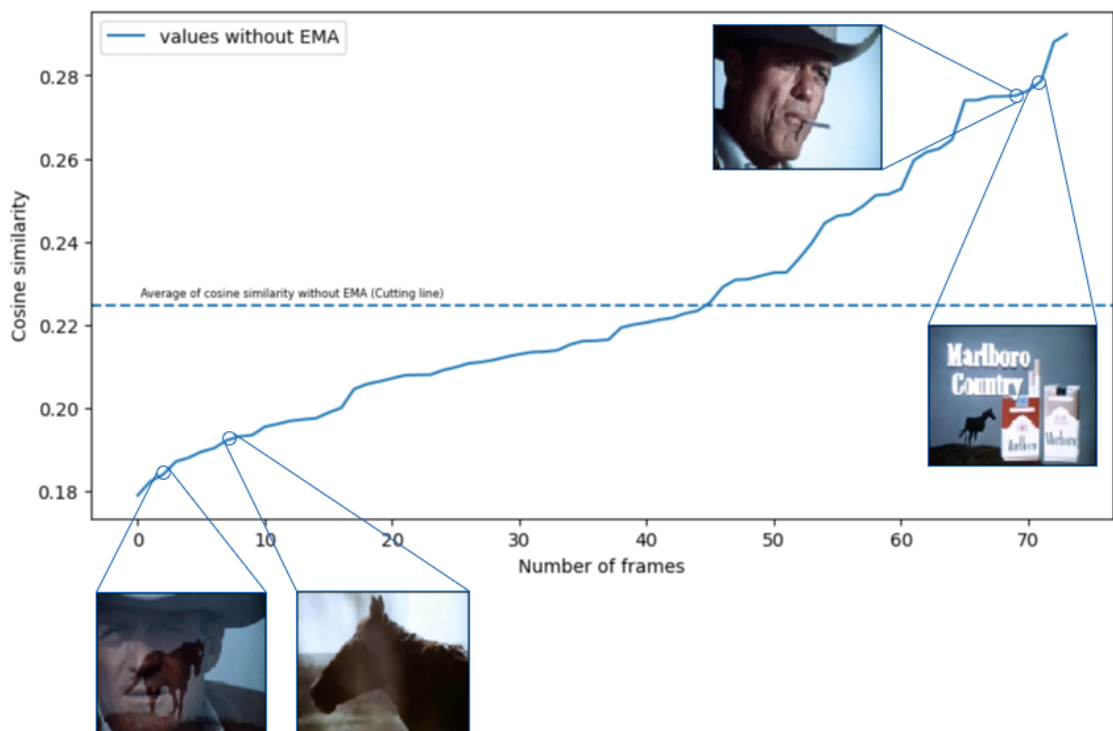


Figure 3: The images are in an orderly manner based on the cosine similarity values.

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

139 2.5 Fine-tuned image classification

140 After filtering the image set with a multimodal model, we applied an image processing model
141 to classify the remaining images further to improve accuracy. Among the publicly available
142 datasets on smoking, we have used the "smoker and non-smoker"³⁸ for augmented³⁹ fine-
143 tuning. We selected the following models for the task. EfficientNet, Inception, ResNet, VGG,
144 and Xception. The EfficientNet B5 version was the best, with an accuracy of 93.75%. Table S1
145 of the supplemental contains our detailed measurement results concerning all models.

146 2.6 Processing of text

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

155 We created blocks of a maximum of 5 elements from the words in our dictionary. Each block
156 contained a random combination of a maximum of 5 words. The blocks are disjoint, so they do
157 not contain the same words. This mixing step was done 10 times. This means that, in one itera-
158 tion, we could form 8 blocks of 5-element disjunct random blocks from our 43-word dictionary.
159 By doing all these 10 times, we produced 80 blocks. However, due to the 10 repetitions, the
160 80 blocks were no longer disjoint. In other words, if we string all the blocks together, we get a
161 dictionary in which every synonym for smoking appears a maximum of 10 times.

162 We made a prompt template to which, by attaching each block, we instructed ChatGPT to gen-
163 erate texts containing the specified expressions. Since ChatGPT uses the Hungarian language
164 well, the generated texts contained our selected words by the rules of the Hungarian language,
165 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

166 We did not specify how long texts should be generated by ChatGPT or that every word of a
167 5-element block should be included in the generated text. When we experimented with Chat-
168 GPT generating fixed-length texts, it failed. Therefore, we have removed the requirement for
169 this.

170 Using this method, we created a smoking-related corpus consisting of 80 paragraphs, 49000
171 characters, and 7160 words. An English example of a generated text is presented in Table
172 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.

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

178 2.7 Human reinforcement

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

190 **3 Results**

191 We collected video materials to test the image processing part of our architecture. The source
192 of the video materials was the video-sharing site YouTube. Taking into account the legal rules
193 regarding the usability of YouTube videos, we have collected 5 pieces short advertising films
194 from the Malboro and Philip Moris companies. We ensured not to download videos longer than
195 2 minutes because longer videos, such as movies, would have required a special approach
196 and additional pre-processing. Furthermore, we downloaded the videos at 240p resolution and
197 divided them into frames by sampling every second. Each frame was transformed to a resolution
198 of 224×224 pixels. We manually annotated all videos. The downloaded videos averaged 64
199 seconds and contained an average of 13 seconds of smoking.

200 With the multimodal filtering technique, we discarded the images that did not contain smoking.
201 Multimodal filtering found 25 seconds of smoking on average in the recording. The accuracy
202 of the identified images was 62%. The multimodal filtering could filter out more than half
203 of the 64-second, on average, videos. We also measured the performance of the fine-tuned
204 EfficientNet B5 model by itself. The model detected an average of 28 seconds of smoking with
205 60% accuracy. We found that the predictions of the two constructions were sufficiently diverse
206 to connect them using the boosting ensemble⁴⁴ solution. By connecting the two models, the
207 average duration of perceived smoking became 12 seconds with 4 seconds on average error
208 and 74% accuracy. The ensemble solution was the best approach since the original videos
209 contained an average of 13 seconds of smoking. We deleted the videos after the measurements
210 and did not use them anywhere for any other purpose.

211 We created training and validation datasets from Hungarian synonyms for smoking using Chat-
212 GPT. We trained our chosen large language models until their accuracy on the validation dataset
213 did not increase for at least 10 epochs. The XLM-RoBERTa model achieved the best perfor-
214 mance on the validation dataset with an F1-score of 96% and 98% accuracy. For the final
215 measurement, we created test data from an online text related to smoking by manual annota-
216 tion.⁴⁵ The text of the entire test data is included in the Table S20 supplemental. The fine-tuned
217 XLM-RoBERTa model achieved 98% accuracy and 0.91 F1 score on the test dataset.

218 **4 Conclusions**

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

229 **Funding**

230 The project no. KDP-2021 has been implemented with the support provided by the Ministry
231 of Culture and Innovation of Hungary from the National Research, Development, and Innova-
232 tion Fund, financed under the C1774095 funding scheme. Also, this work was partly funded
233 by the project GINOP-2.3.2-15-2016-00005 supported by the European Union, co-financed by
234 the European Social Fund, and by the project TKP2021-NKTA-34, implemented with the sup-
235 port provided by the National Research, Development, and Innovation Fund of Hungary under
236 the TKP2021-NKTA funding scheme. In addition, the study received further funding from the
237 National Research, Development and Innovation Office of Hungary grant (RRF-2.3.1-21-2022-
238 00006, Data-Driven Health Division of National Laboratory for Health Security).

239 **References**

- 240 [1] for Economic Co-operation, O.; Development Daily smokers (indicator). 2023.
- 241 [2] Organization, W. H. Tobacco. 2022.
- 242 [3] Chapman, S.; Davis, R. M. *Tobacco Control* **1997**, 6, 269–271.
- 243 [4] Pechmann, C.; Shih, C. *Irvine, California: Graduate School of Management, University of*
244 *California, Irvine* **1996**,

- 245 [5] Kong, G.; Schott, A. S.; Lee, J.; Dashtian, H.; Murthy, D. *Tobacco Control* **2022**,
- 246 [6] Fielding, R.; Chee, Y.; Choi, K.; Chu, T.; Kato, K.; Lam, S.; Sin, K.; Tang, K.; Wong, H.; Wong, K.
247 *Journal of Public Health* **2004**, 26, 24–30.
- 248 [7] Fu, R.; Kundu, A.; Mitsakakis, N.; Elton-Marshall, T.; Wang, W.; Hill, S.; Bondy, S. J.; Hamil-
249 ton, H.; Selby, P.; Schwartz, R.; others *Tobacco Control* **2023**, 32, 99–109.
- 250 [8] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings
251 of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.
- 252 [9] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchi-
253 cal image database. 2009 IEEE conference on computer vision and pattern recognition.
254 2009; pp 248–255.
- 255 [10] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *arXiv preprint arXiv:1810.04805* **2018**,
- 256 [11] Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning
257 books and movies: Towards story-like visual explanations by watching movies and reading
258 books. Proceedings of the IEEE international conference on computer vision. 2015; pp
259 19–27.
- 260 [12] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; others Improving language under-
261 standing by generative pre-training. 2018.
- 262 [13] Common Crawl. 2022; Accessed: 2022-06-01.
- 263 [14] Blei, D. M.; Ng, A. Y.; Jordan, M. I. *Journal of machine Learning research* **2003**, 3, 993–
264 1022.
- 265 [15] Pennington, J.; Socher, R.; Manning, C. D. GloVe: Global Vectors for Word Representation.
266 Empirical Methods in Natural Language Processing (EMNLP). 2014; pp 1532–1543.
- 267 [16] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. *Advances in neural information*
268 *processing systems* **2013**, 26.
- 269 [17] Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. *Transactions of the association for compu-*
270 *tational linguistics* **2017**, 5, 135–146.
- 271 [18] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:1908.10084* **2019**,

- 272 [19] Clark, J. H.; Garrette, D.; Turc, I.; Wieting, J. *Transactions of the Association for Computa-*
273 *tional Linguistics* **2022**, *10*, 73–91.
- 274 [20] Arthur, D.; Vassilvitskii, S. *k-means++: The advantages of careful seeding*; 2006.
- 275 [21] Ali, S.; Masood, K.; Riaz, A.; Saud, A. Named Entity Recognition using Deep Learning: A
276 Review. 2022 International Conference on Business Analytics for Technology and Security
277 (ICBATS). 2022; pp 1–7.
- 278 [22] Simonyan, K.; Zisserman, A. *arXiv preprint arXiv:1409.1556* **2014**,
- 279 [23] Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proceedings
280 of the IEEE conference on computer vision and pattern recognition. 2017; pp 1251–1258.
- 281 [24] Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks.
282 International conference on machine learning. 2019; pp 6105–6114.
- 283 [25] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception archi-
284 tecture for computer vision. Proceedings of the IEEE conference on computer vision and
285 pattern recognition. 2016; pp 2818–2826.
- 286 [26] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time
287 object detection. Proceedings of the IEEE conference on computer vision and pattern
288 recognition. 2016; pp 779–788.
- 289 [27] Lin, J.; Chen, Y.; Pan, R.; Cao, T.; Cai, J.; Yu, D.; Chi, X.; Cernava, T.; Zhang, X.; Chen, X.
290 *Computers and Electronics in Agriculture* **2022**, *202*, 107390.
- 291 [28] Liu, Y.; Guo, Y.; Liu, L.; Bakker, E. M.; Lew, M. S. *Pattern Recognition* **2019**, *93*, 365–379.
- 292 [29] Liu, Z.; Chen, F.; Xu, J.; Pei, W.; Lu, G. *IEEE Transactions on Circuits and Systems for Video*
293 *Technology* **2022**,
- 294 [30] Rao, A.; Xu, L.; Xiong, Y.; Xu, G.; Huang, Q.; Zhou, B.; Lin, D. A local-to-global approach
295 to multi-modal movie scene segmentation. Proceedings of the IEEE/CVF Conference on
296 Computer Vision and Pattern Recognition. 2020; pp 10146–10155.
- 297 [31] Gagliardi, A.; de Gioia, F.; Saponara, S. *Journal of Real-Time Image Processing* **2021**, *18*,
298 2085–2095.

- 299 [32] Bianco, F.; Moffett, C.; Abunku, P.; Chaturvedi, I.; Chen, G.; Dobler, G.; Sobolevsky, S.; Kirchner,
300 ner, T.; others *Authorea Preprints* **2021**,
- 301 [33] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:2004.09813* **2020**,
- 302 [34] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.;
303 Mishkin, P.; Clark, J.; others Learning transferable visual models from natural language
304 supervision. International conference on machine learning. 2021; pp 8748–8763.
- 305 [35] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.;
306 Ott, M.; Zettlemoyer, L.; Stoyanov, V. *CoRR* **2019**, *abs/1911.02116*.
- 307 [36] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; De-
308 hghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others *arXiv preprint arXiv:2010.11929* **2020**,
- 309 [37] Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. *ArXiv* **2019**, *abs/1910.01108*.
- 310 [38] Khan, A. *Mendeley Data* **2020**, *1*.
- 311 [39] Shorten, C.; Khoshgoftaar, T. M. *Journal of big data* **2019**, *6*, 1–48.
- 312 [40] Viola, T. *Ellentétes jelentésű szavak adatbázisa*; Tinta Könyvkiadó, 2012.
- 313 [41] Nemeskey, D. M. Introducing huBERT. XVII. Magyar Számítógépes Nyelvészeti Konferencia
314 (MSZNY2021). Szeged, 2021; p TBA.
- 315 [42] Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. *CoRR* **2018**, *abs/1810.04805*.
- 316 [43] Reimers, N.; Gurevych, I. *arXiv preprint arXiv:1908.10084* **2019**,
- 317 [44] Dietterich, T. G. Ensemble methods in machine learning. Multiple Classifier Systems: First
318 International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1. 2000;
319 pp 1–15.
- 320 [45] Center, H. P. *Egészség Elvitelre*. 2023.