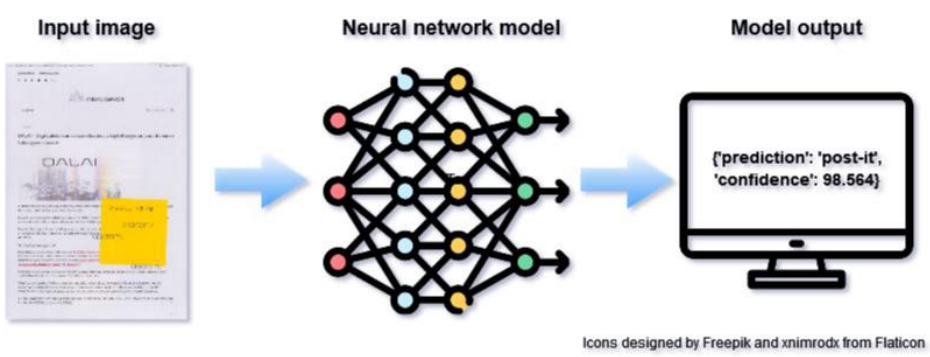
Post-it note identification component

Background, benefits and purpose of use of the component

Documents on which one or more post-it notes have been attached may be scanned in the document digitisation process. If the location of the post-it note affects the readability of the document's content, it is useful if such cases can be identified and, if necessary, scanned again. Machine identification of post-it notes thus helps improve the quality of digital material and reduces the need for manual quality control.

What does the component do?



The component processes the image files received as input one at a time and forms a prediction of the content of the image based on the parameters it has learned during the training. If the component identifies a post-it note in the image with a probability of more than 50%, the image is classified as incorrect. The component returns the predicted class to the user.

How is the component used?

The component is available both as part of the Arkkiivi user interface (http://www.arkkiivi.fi/) and as a standalone application offering an application programming interface (API) on the DALAI project's GitHub page (https://github.com/DALAI-project/PostitAPI). More detailed instructions on how to use the component in different environments can be found on the attached website.

Component training

Image files in .jpg format aligned to 224 x 224 pixels have been used for component training. A total of 55,000 documents have been used in the training, of which approximately 4,000 are documents containing post-it notes. Images containing post-it notes have been both self-made and extracted manually from images classified as incorrect in connection with the mass digitisation process of the National Archives. The documents contain a variable number of post-it notes of different size and colour, placed in different parts of the document, and the notes may also contain text.

Although efforts have been made to compile a variety of example cases for the training material, the amount of the material is limited and errors also occur in the classification of the component. Among others, content elements that resemble post-it notes, such as coloured rectangular text fields, may lead to incorrect classification.