The Interactive Classification System

by Andrea Esuli (ISTI-CNR)

ISTI-CNR released a new web application for the manual and automatic classification of documents. Human annotators collaboratively label documents with machine learning algorithms that learn from annotators' actions and support the activity with classification suggestions. The platform supports the early stages of document labelling, with the ability to change the classification scheme on the go and to reuse and adapt existing classifiers.

The Interactive Classification System (ICS) [1] is a novel open-source software [L1] from ISTI-CNR that is designed to support the activity of manual text classification. The application uses machine learning to continuously fit automatic classification models that are in turn used to actively support its users with classification suggestions, or to produce automatic classification of large datasets.

The aim of ICS is to support the activity of manual text classification, in collaborative scenarios, with a special attention to the early stages of labelling, i.e., when the dataset may be not yet complete, the classification schema is not yet consolidated. These scenarios receive less attention from the research on machine learning methods for text classification, which typically assume a more solid setup on which to operate.

ICS can be deployed as a web application (see Figure 1) and multiple users can collaboratively build datasets, define classification schemas, and label the documents in the datasets according to any of the classification schemas. A unique feature of ICS is that it gives its users total freedom of action: they can at any time modify any classification schema, any dataset, and any label assignment, possibly reusing any relevant information from previous activities. During these activities a machine-learning agent monitors those events and keeps a pool of automatic classifiers up to date.

Such freedom of action has an impact on how the machine learning algorithms used in the system must operate. The machine-learning approach used for ICS can be defined as "unobtrusive machine learning", as it never actively requests any action from the users, while it instead silently observes their actions, continuously adapting the automatic classification mod-

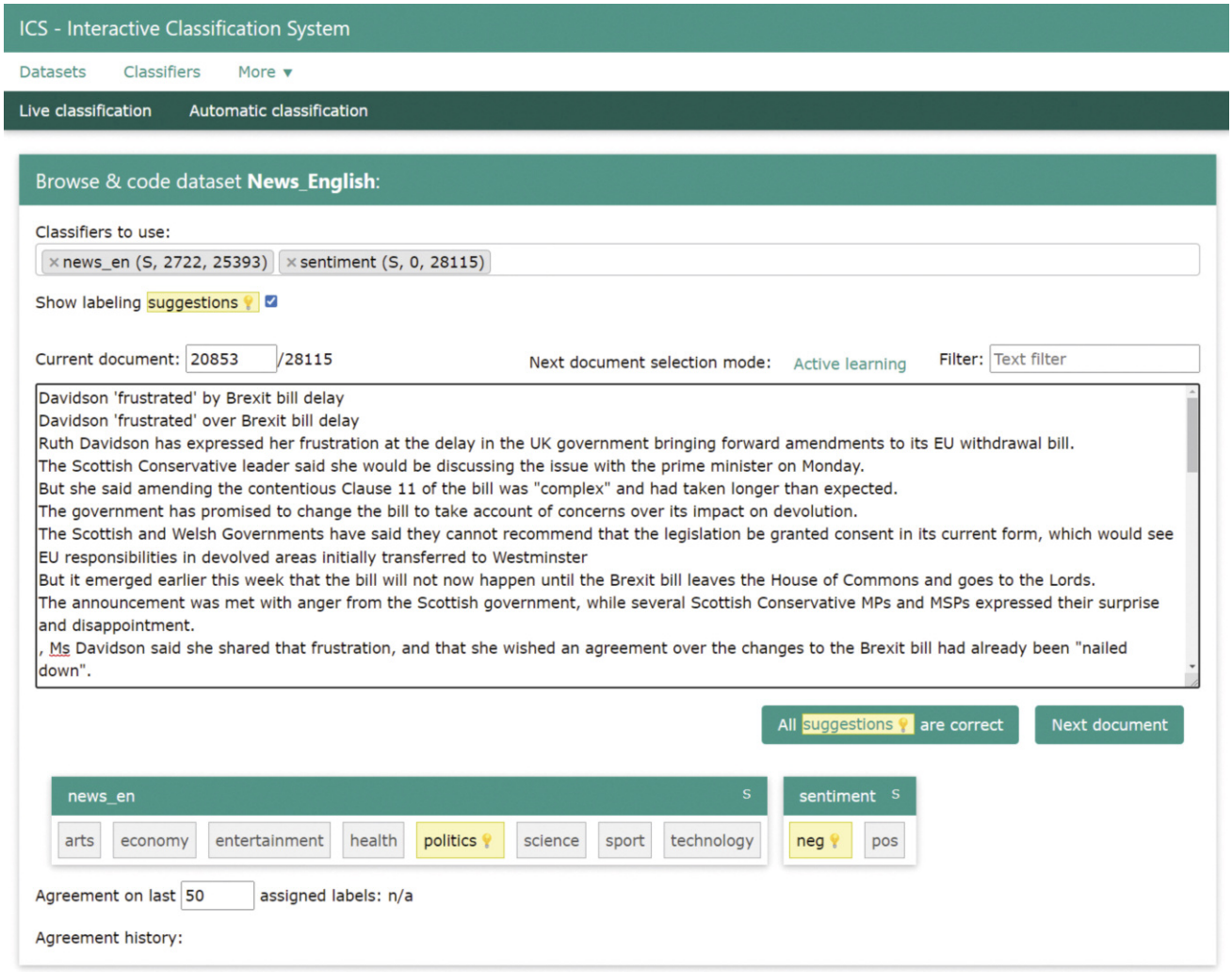

Figure 1: A screenshot of ICS. A document is shown, with two classifiers selected for labelling, i.e., by topic and by sentiment. Labels highlighted in yellow are the ones suggested by the machine learning model that ICS continuously updates according to users' actions.

els that are in turn used to provide the classification suggestions.

The unobtrusive approach challenges many of the assumptions made by the typical machine-learning approaches, i.e., batch-learning-based approaches, which require a complete training set before being able to be used, or active-learning-based approaches, which consider the user as a human-in-the-loop that acts on request from the algorithm, with very limited freedom of action.

The implementation of the unobtrusive machine learning required by ICS uses an online-learning algorithm (Passive-Aggressive, [2]). This algorithm is not yet sufficient to produce a usable implementation, as the algorithm can work in an online fashion only if the vector space in which the text documents are projected does not change. This is not the case for ICS, as its document datasets can be modified. A simple case of this is labelling a dataset of tweets that is continuously fed by an active streaming query. ICS solves this issue by using Lightweight Random Indexing [2], which defines a fixed vector space that is independent of any feature extracted from text, and it is efficient in reducing the number of dimensions in the vector space while not losing information from the resulting vectorial representation of text. This efficiency is a key aspect in having fast update time for any classifier, both due to lower computational complexity and due to reduced I/O from the DB that stores all the classification models. Many other theoretical and technological "tricks" are used in ICS to provide its users with responsive and accurate automatic classification, e.g., feature hashing, variable mini-batch learning, mini-sample-based active learning. The paper that presents ICS in details [1], also reports on some experimental evaluation that compares the performance of the model against traditional batch-learning approaches, including evaluating the possibility of performing transfer learning, reusing an existing classifier.

ICS is distributed under the BSD license and can be easily installed using the pip Python package installer (ics-pkg). More information, including installation and usage video tutorials are available on the GitHub repository [L1].

Link:
[L1] https://github.com/aesuli/ics

References:

[1] A. Esuli, "ICS: Total Freedom in Manual Text Classification Supported by Unobtrusive Machine Learning", in IEEE Access, vol. 10, pp. 64741-64760, 2022. https://doi.org/10.1109/access.2022.3184009
[2] K. Crammer, et al., "Online Passive-Aggressive Algorithms", in Journal of Machine Learning Research, 7, 551-585, 2006.
[3] A. Moreo, A. Esuli and F. Sebastiani, "Lightweight random indexing for polylingual text classification", in Journal of Artificial Intelligence Research 57, pp. 151-185, 2016. https://doi.org/10.1613/jair.5194

Please contact:

Andrea Esuli, ISTI-CNR, Italy
andrea.esuli@isti.cnr.it

How Quickly do Trees Grow?

by Refiz Duro (Austrian Institute of Technology), Hanns Kirchmeir (E.C.O. Institut für Ökologie Jungmeier), Anita Zolles (Bundesforschungs- und Ausbildungszentrum für Wald, Naturgefahren und Landschaft) and Günther Bronner (Umweltdata)

Changing climatic circumstances have a significant impact on forests: besides higher temperatures, more intense and frequent storms and drought spells affect forest growth. To see what the future is bringing, and to be able to deal with forest conservation and management, it is necessary to answer the question "how quickly do trees grow in an environment of climate change?" We take on a challenge to answer this question by integrating state-of-the-art data collection and AI-based methods.

Forests are the largest terrestrial sinks for carbon and some of the richest biological areas on Earth. Climate change is, however, affecting forests through increasing temperatures, changing precipitation patterns and the growing number of biotic and abiotic disturbances [1]. Saving forest ecosystems is thus one of the key measures to mitigate climate change and save biodiversity. To maintain and improve forest biodiversity and forest resilience to climate change, updated forest policies and forest management strategies are being developed and implemented in adaptive forest management. They all require up-to-date data of the forest conditions including the vitality and health of trees, as well as the ongoing tree growth (i.e., carbon sequestration).

Initially, tree and forest growth have been assessed mainly for economic reasons in order to build forest yield tables as simple "growth models" and a basis for improved forest management and taxation. Only within the past 50 years, improved tree and forest growth models have become available (e.g., [2]). They do not, however, allow for consideration of instantaneous changes in growth due to climatic extremes (e.g., drought, heat) and the changes of phenological patterns (i.e., length vegetation season).

Nowadays, new and more accurate measurement equipment has become available and new tree and forest characteristics have become the focus of forest and environmental science. The latest measurement equipment allows, for example, to assess intraday variation of tree radial growth (automatic dendrometers), to assess the water status of trees during a day (sap flow). This is used to characterise the tree and crown architecture including their habitat and microhabitat functions with terrestrial laser scanning in highest precision, or to assess tree health and vitality with airborne and satellite imaging / laser scanning. These new measurement techniques provide a huge amount of quantitative forest data, but their correct analysis and interpretation strongly requires advanced analytics to fully utilise the obtained data, achievable nowadays with the advanced probabilistic and machine learning approaches (e.g., neural networks).