Introducing a transparency-driven platform for creating, comparing and explaining machine learning pipelines

Tom Klaver*[1], Erik Tjong Kim Sang[1], Aysenur Bilgin[2], Kim Smeenk[3],
Laura Hollink[2], Jacco van Ossenbruggen[2], Frank Harbers[3], and Marcel Broersma[3]
[1]*Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam*
[2]*Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam*
[3]*University of Groningen, 9712 CP Groningen*

***Abstract*—Machine learning offers great promises for many tasks relevant to science and industry. Before users accept machine-generated solutions, they need to be convinced of the quality of the system. However, the internal operation of a machine learning system is often complicated. In this demonstration, we present a transparency-driven platform that can create, compare and explain machine learning pipelines by means of graphical and tabular visualizations. The platform displays a genre classification task in digital newspaper archives, however it can be employed for any other machine learning task.**

***Index Terms*—Machine learning, transparency, genre classification, digital humanities**

I. Introduction

Increasing amounts of data have facilitated the development of data-driven approaches that rely on complex machine learning algorithms. With respect to extracting meaningful insights, large amounts of data usage for automated decision making has been developing concerns [3] [7]. Inspired by the initiatives such as the Responsible Data Science (RDS[1]) and Explainable AI (XAI) [4] that aim to tackle these concerns by increasing awareness on transparency, we have built a web-based platform to encourage the trusted use of machine learning pipelines. The demonstrated application targets not only data scientists (as in many existing architectures such as TensorBoard[2], Spark ML[3] and Skater [2]) but also lay users and can be applied to any other machine learning task.

In this demonstration, we present the functionality of the platform using a genre classification task (for further details about this research, see [1]). We address the following questions, which are central to most of the existing machine learning tasks from both developer and lay user perspectives:

1) How can we optimize the machine learning pipelines for task-specific accuracy in a user-friendly way?
2) What form of representations facilitate communicating the important criteria driving the automated decision making process? In particular, a user should be able to understand the type of errors being made and gauge to what extent the model can be trusted for a given task.
3) Can we provide a user-friendly interface that allows a lay user to utilize complex machine learning pipelines by uploading data sets, comparing pipelines and visualizing the results as well as allowing them to understand the decisions given by the pipelines?

In the next section, we provide details regarding the live actions and interactions offered by the platform.

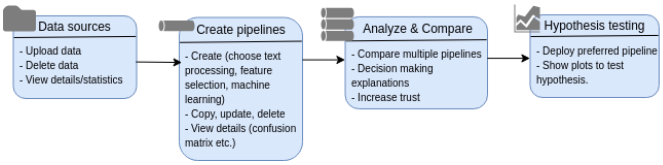

Fig. 1. Major platform functionality

II. Demonstration of the platform

The platform[4] is developed using Flask[5], MongoDB[6], scikit-learn [5], Frog[7] natural language processing suite and LIME [6]. It offers two interfaces: 1) web-based graphical user interface (GUI) and 2) Jupyter Lab[8] interface. Both interfaces allow for uploading data sets, constructing custom pipelines, analyzing results of individual pipelines as well as comparing multiple pipelines, providing explanations on local and global levels and deploying the pipelines to test hypotheses.

As summarized in Fig. 1, the web-based GUI has the following page views for demonstration:

1) Data view (Fig. 2): Users can upload, view and delete data sets. Clicking on a data set shows the task-specific statistics of the data set (e.g. size of the data set, distribution of features, etc.).
2) Pipelines view: Users can customize pipelines by choosing from a set of predefined text processing and feature selection steps as well as by configuring the desired machine learning algorithm (currently a limited number

* E: t.klaver@esciencecenter.nl. T: +31651894359
[1]http://www.responsibledatascience.org/
[2]https://www.tensorflow.org/guide/summaries_and_tensorboard
[3]https://spark.apache.org/docs/1.2.2/ml-guide.html
[4]The code is available at: https://github.com/newsgac/platform
[5]http://flask.pocoo.org/
[6]https://www.mongodb.com
[7]https://languagemachines.github.io/frog/
[8]https://jupyterlab.readthedocs.io/en/stable/

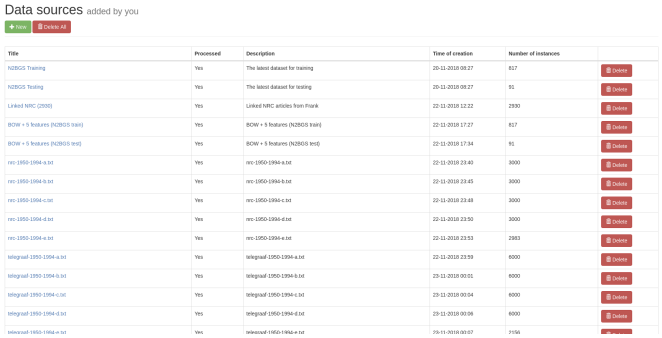

Fig. 2. Data view: users can upload, view and delete data sets. https://bit.ly/2AN7cZs

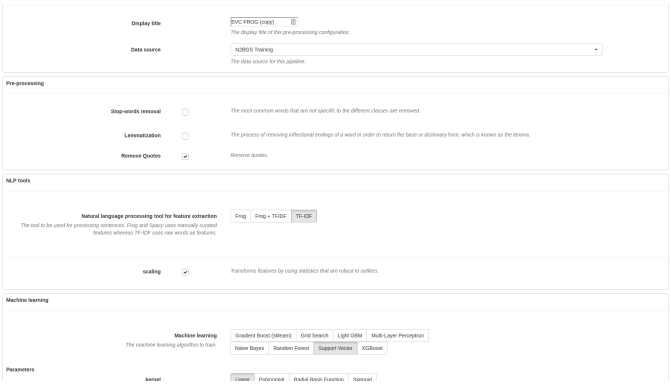

Fig. 3. Pipeline view: Users can customize pipelines by choosing from a set of predefined text processing and feature selection steps as well as by configuring the desired machine learning algorithm. https://bit.ly/2sq37FZ

of classifiers from scikit-learn [5] library are implemented) (Fig. 3). The option to use optimized algorithm parameters by means of grid search[9] is also available to the users. Once a pipeline is constructed, users can track the training status and view the run-times of training upon completion. Once successfully trained, pipelines can be copied, analyzed and deleted through the interface. The analysis view allows the users to examine several performance metrics tailored for the task (e.g. accuracy, recall, confusion matrix, etc. for classification tasks) as well as global-level explanations for the pipeline depending on the type of the machine learning algorithm selected (e.g. feature importance ranking plots). The trained pipelines are stored in database and can be deployed on other data sets.

3) ACE (Analyze-Compare-Explain) view (Fig. 4): Users can critically analyze and understand the decision making processes of not only a single pipeline but multiple pipelines. For this purpose, we make us of several graphical views along with tabular views that allow the user either explore the local-level explanation concerning one instance (i.e. one data point) or determine the majority vote (agreed prediction) of multiple pipelines. Using this

[9]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

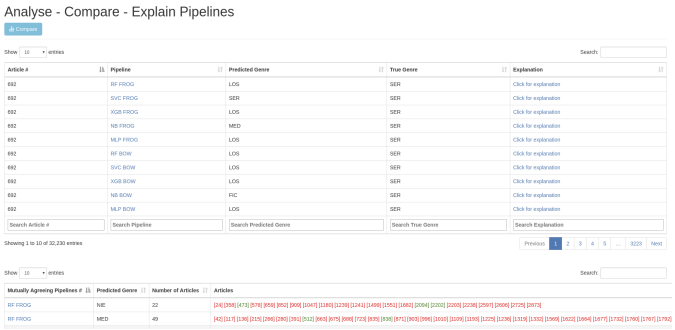

Fig. 4. ACE view: Users can critically analyze and understand the decision making processes of multiple pipelines. https://bit.ly/2FwaDa8

view, the users can gather meaningful insights regarding the decision making processes of the complex machine learning pipelines and develop trust in them accordingly.

4) Hypothesis testing: Users can deploy the trusted pipeline in other data sets for the purpose of retrieving statistics that can help them test their hypotheses. In this view, the platform displays plots for a visual analysis of the predictions.

III. Summary

We present a ready-to-use platform for creating, comparing and explaining machine learning pipelines and demonstrate the practical impact of transparency for a genre classification task.

Acknowledgments

The study described in this paper was executed as a part of the NEWSGAC project which is funded by the Netherlands eScience Center[10] and CLARIAH[11].

References

[1] A. Bilgin, E. Tjong Kim Sang, K. Smeenk, L. Hollink, J. van Ossenbruggen, F. Harbers, and M. Broersma. Utilizing a transparency-driven environment toward trusted automatic genre classification: A case study in journalism history. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 486–496, Oct 2018.

[2] Pramit Choudhary, Aaron Kramer, and contributors datascience.com team. Skater: Model Interpretation Library, March 2018.

[3] Natasha Duarte, Emma Llanso, and Anna Loup. Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*. PMLR, 2018.

[4] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[7] Wil M. P. van der Aalst, Martin Bichler, and Armin Heinzl. Responsible data science. *Business & Information Systems Engineering*, 59(5):311–313, Oct 2017.

[10]escience center.nl

[11]clariah.nl