A prototype machine learning and data visualization platform for text classification

Joseph Annis
Northeastern University
Boston, MA, USA
annis.jo@husky.neu.edu

Dhruv Sheth
Northeastern University
Boston, MA, USA
sheth.dhr@husky.neu.edu

Aleszu Bajak
Northeastern University
Boston, MA, USA
a.bajak@neu.edu

Abstract

Machine learning algorithms are being used by journalists around the world to enhance virtually every step of the news production process.[1] Indeed, machine learning techniques have been used by news outlets like The Marshall Project to cluster cities by crime patterns, The New York Times to predict advertising engagement, BuzzFeed News to predict surveillance aircraft flight patterns, and The Washington Post and Bloomberg to classify political rhetoric on social media.[2] [3] [4] But to date these machine learning techniques remain in the domain of experts. To do any amount of significant analysis and useful visualization, one needs a relatively high level of technical skill with programming languages such as R or Python, an understanding of how machine learning models work and how to interpret them, and the ability to harvest vast amounts of data.

We propose and have designed a platform where journalists can find and implement existing machine learning models specifically on text classification problems, view and analyze the data associated with a given model, create rich visualizations of that data, and contribute datasets and models of their own.

Our beta prototype web application, built with React and Typescript, aims to attach an easy-to-use interface and instructional information to technical elements that would normally require knowledge of programming and advanced technical concepts. Our use-case imagines a scenario where a journalist wishes to perform a text classification task on candidate tweets for a story on the 2020 presidential election to understand the topics being discussed in that corpus of tweets -- i.e. how much a candidate is messaging on climate change or immigration. Using our platform, the journalist 1) uploads a spreadsheet of thousands of candidate tweets; 2) searches for and implements a topic modeling algorithm; and 3) creates visualizations to inform their article as well as a statistical output to cite as hard evidence.

This drag-and-drop machine learning and data visualization prototype 1) would allow users who have little to no knowledge of machine learning to understand how ML can be applied; 2) is complex enough such that experienced programmers can still wield it for journalism; 3) would accept CSVs or makes calls to an API through a visual interface, visualize large quantities of data effectively; 4) would have a robust yet easy-to-use data visualization builder; and 5) would allow collaborators to find and discuss projects easily.

This prototype platform aims to make machine learning techniques more accessible to newsrooms. We were heavily inspired by Overview, Jonathan Stray's open-source document analysis system, and OpenML, an "open science platform for machine learning," as well as the aforementioned classification tasks performed by The Washington Post and Bloomberg.[5] [6]

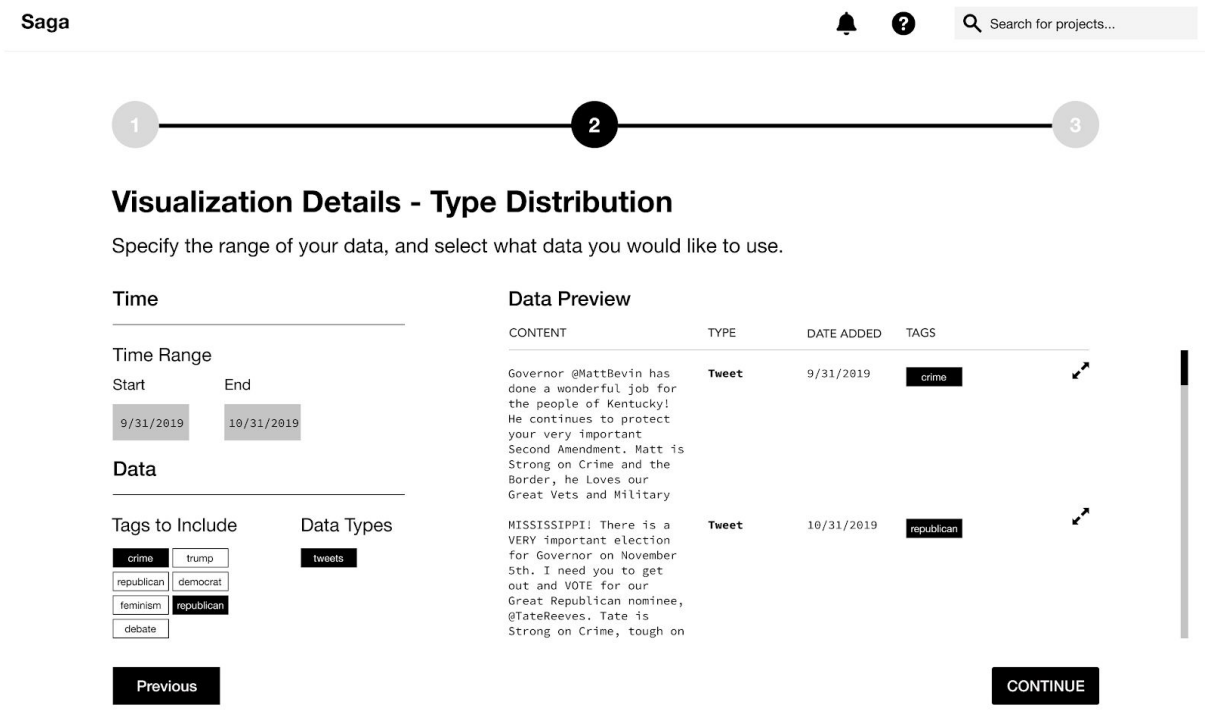

Figure 1: Once the dataset is classified, users can select and visualize the topics they're interested in.

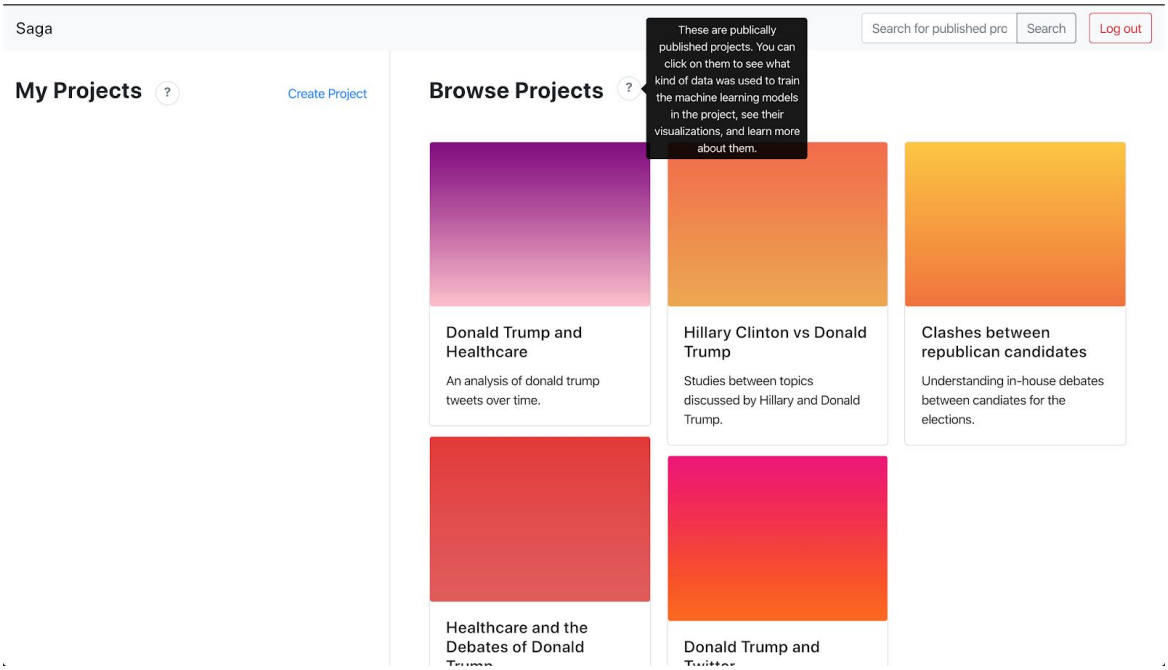

Figure 2: Our prototype allows users to contribute and search through a database of publicly-published projects.

References

[1] Diakopoulos, N. (2019). "Automating the News: How Algorithms Are Rewriting the Media." Harvard University Press. ISBN: 978-0674976986.
[2] Ivancsics, B & Hansen, M. (2019). Columbia Journalism Review. "Actually, it's about Ethics, AI, and Journalism: Reporting on and with Computation and Data." Retrieved from https://www.cjr.org/tow_center_reports/ai-ethics-journalism-and-computation-ibm-new-york-times.php
[3] Schaul, K & Uhrmacher, K. (2019). The Washington Post. "The issues 2020 Democrats are running on, according to their social media." Retrieved from https://www.washingtonpost.com/graphics/politics/policy-2020/priorities-issues/
[4] McCartney, A. (2019). Bloomberg. "What the Democratic Presidential Candidates Care About, in 44,000 Tweets." Retrieved from https://www.bloomberg.com/graphics/2020-democratic-presidential-candidate-policies/#methodology
[5] Overview. Retrieved from https://www.overviewdocs.com/
[6] OpenML. Retrieved from https://www.openml.org/