Andrew Nieto, Jacob Sandy, Kyle Laycock, and Christian D'Amico
Dr. Philippe J. Giabbanelli
Ms. Alia Levar Wegner
LAURE Award Reflective Essay

The Application of Machine Learning to Digitized Postcards as Artifacts

When machine learning is mentioned in the news, it is rarely in the context of historical artifacts. Often, the public learns about what machine learning promises through news about optimizing workflows, improving trading algorithms, or even changing health outcomes for patients. However, this project demonstrates that machine learning techniques and tools can be used to develop new insights and collect new information on historical artifacts such as the Bowden Postcard Collection at Miami University.

When we were first introduced to this topic, it was solely within the confines of the classroom; a few members of the group were not even aware that there was a special collection of postcards here at the library, much less that it had been digitized. When our group came together, we first met with the digital collections librarian Alia Wegner to discuss the digital collection and ask questions about it, such as its origin, purpose, the process of digitizing the postcards, possible areas of research, and more. Soon after, we were introduced to the dataset in question: an excel spreadsheet of over 20,000 rows, each with nearly 39 attributes. Many of those attributes proved to be not useful or purely informational (e.g. the copyright, the digital publisher), but some of these attributes were potentially holding a trove of insights to be made about the dataset (e.g. the subject of the picture, the message written on the postcard, etc.). Once we had the digitized dataset, we set about the next stage of our process: examining and cleaning the dataset.

One of the most important parts of machine learning is coming to a greater understanding of what your data is before you decide what you can do with it. Unfortunately, many datasets come to researchers and teams messy and maybe even incomplete, so it is up to the researchers to determine what best practices should take place to ensure the data is "cleaned", or made easier to use for machine learning purposes. Once we had completed that task it was then onto the main stage of applying machine learning

algorithms and libraries (i.e., a set of tools used by software developers to write their programs) to the dataset.

What we first noticed when examining the dataset was the variety of possibilities that we could explore with this dataset. After some consideration, over the next two milestones (i.e., smaller deadlines) we divided up the tasks amongst ourselves to generally adhere to several main categories. We picked categories that we felt could give us the most useful information. The first category of tasks that were explored throughout this project was sentiment analysis or the judging of the general mood of the author/writing of a text on a range from negative to positive. The second category that was exploring the pictures on the front of the postcards, whether that was color identification of the main color palettes used on the front of the cards, to object detection where we attempted to identify what the postcard was depicting. The third was analyzing the subjects depicted on the postcards in relation to time and space to attempt to see if there was any major difference between the kinds of postcards people were sending based on location or when they were sending it in history. Finally, the fourth category was topic modeling, or the analysis of the text of cards to main points people wrote about in their cards (i.e. Someone wrote about an upcoming visit with the person they are writing to). The main programming language we worked in was Python (Also the near de facto data analytics and machine learning language) and we utilized a variety of different machine learning libraries to accomplish our tasks (e.g. Gensim for topic modeling, Vader for sentiment analysis, etc.).

Several challenges were encounter by our team members. Whether that be libraries that would not install properly on someone's computer or unexpected outcomes, programming is an exercise in will and patience as much as it is in knowledge and skill. Some challenges included running very intense computational models/algorithms on our computer that could take long periods of time to complete learning all new libraries, and packages we had never encountered before in a very short amount of time. Because of this, we learned how to learn quickly on the job, something commonly seen in a professional work environment. One of the more enlightening challenges that we encountered in the project was

centered around color detection of the primary colors of the postcards. The librarian discussed with us the many possibilities that come with having an addition to the dataset in which one can search by color for a particular art class they are taking, or perhaps a color-blind student who can better find the digitization of the color prints more accessible if they were labeled with a primary color or two. In response to this, our team created a color identification UI that would communicate with the dataset to bring the user several postcards that had similar colors to one the user would enter. It is not often that students can see their work be used in ways that can benefit other students or researchers, and the team is pleased that we have at least built a tool that can do that.

By the end of this project, the team had gained not only a better understanding of machine learning libraries, techniques, and processes but also actively utilized library resources and collections to gain a deeper understanding of a collection of artifacts that had been digitized for a larger audience. Additionally, our team was able to actively work with the library staff on creating a tool that could be used by researchers and other students for their projects, especially if they had certain physical limitations such as color blindness. Overall, it has been a pleasure for our team to work with the library's staff and resources on such an amazing project. We hope that the collection and tool continue to grow for future students, researchers, and staff to explore.