

Jamie Reintjes

As part of my major in International Economics and Chinese, I was nominated to complete the Senior Honors Thesis. While brainstorming on potential research topics, I was inspired by a recent uptick in media attention being devoted to China's role in the U.S. fentanyl epidemic, namely as fentanyl's number one manufacturer. I was interested in capturing American's sentiment towards China, but I was unfamiliar with how to measure such a concept. Fortunately, Eric Lease Morgan, Digital Initiatives Librarian of the Navari Family Center for Digital Scholarship (CDS), was able to provide both a space to explore different models and methods used for conducting sentiment analysis, as well as assist me in collecting and curating a news media database. Utilizing his skills in text mining and Natural Language Processing (NLP), Eric collected over 14,000 articles that matched our key search terms of "China" and "fentanyl," each with a specific sentiment score that I was able to use in my final thesis paper wherein I analyzed the causal relationship between fentanyl mortality and news media coverage of China and fentanyl.

In our first meeting, I approached Eric with these initial concepts before we explored the different types of media we could potentially analyze. I was aware of political advertising, news, and social media posting about China's role as the primary manufacturer of fentanyl products and precursors, but my thesis advisor had no experience in measuring sentiment. This was Eric's specialty, having worked with him in the past on other data collection projects. We explored a variety of potential data sources, such as mining tweets using the Twitter Streaming API, but ultimately, we decided on using Nexis Uni as our primary database for news-related media. Nexis Uni is an online database and university research tool that we were able to access through the Hesburgh Library's subscription to their services. They provide a digital database of full-text newspaper articles submitted by the publications themselves. Given their formatting and Eric's familiarity with the database, it proved an adequate tool to use to collect newspaper articles about fentanyl and China.

Utilizing his skills in NLP, Eric then designed a Python program that would download all U.S. news articles from 2012 to 2020 that were about China and fentanyl. With this program, we were able to capture all relevant articles that were specifically about China and fentanyl, as many articles may mention the words China and fentanyl, but could be referring to other themes or be focused on a different topic altogether. After parsing the raw text, Eric was then able to employ a topic modeling operation and present certain descriptive statistics about the sample. Topic modeling is an unsupervised machine learning process used to innumerate latent themes in any corpora. Specifically, Eric executed a latent Dirichlet allocation (LDA) strategy, which is a specific Bayesian network or statistical probability framework that allows us to group together certain words or word phrases into specific categories or thematic boxes. Spending time in the library, working with Eric in the CDS, gave me a greater appreciation for the library's resources. Our collaboration created both a physical and digital hub of our combined research, and Eric empowered me to create new knowledge in a digital environment.

From the Hesburgh Library's subscription to Nexis Uni, Eric and I curated 5,550 articles published in the U.S. from 2012 to 2020 that were about China and fentanyl. From these 5,550 articles, I was capable of identifying state-level location data, and combined with the year each article was published, I created a panel set of 123 state-year observations, thus providing a unique article count per observation and the corresponding average sentiment score. This way, I

was able to employ a more traditional Difference-in-Differences (DiD) regression model, which analyzes the impact the change in fentanyl mortality affects the changes in media coverage and media sentiment towards China's relationship with fentanyl manufacturing and trafficking. In order to access this data, I used a program called STATA, which Hesburgh provides on all in-house desktop computers. I am grateful for the school's subscription to STATA, as it was otherwise too expensive for me to purchase individually.

Throughout this process, Eric would introduce me to new technologies and resources the library provided, thus inspiring me to continue this research. One area in particular that I wish to further explore is improvements to the model we used for sentiment analysis. Eric introduced me to a readily available Python library called TextBlob, which parsed out sentences that included our keyword "China," to create sentiment scores. TextBlob employs a dictionary-style approach using a list of words with their corresponding form, sense, polarity, subjectivity, and intensity, among other variables to generate a sentence's sentiment score. Given time constraints, we did not have the capability to create our own model; instead, we used the pre-existing model that was created using movie reviews. As a result, my findings were generally inconclusive due to some incongruencies in the sentiment scoring of the text data.

However, thanks to my experience with Eric and the Hesburgh Library, I am excited to continue expanding on this research in graduate school. This opportunity allows me to take the skills I have gained from engaging with the Hesburgh Library services, especially those gained in collaboration with Digital Initiatives Librarian Eric Morgan. Thanks to the library's vast network of subscriptions and database services, I have gained better familiarity with the sources that are readily available for text mining and data curation. Thanks to the discussions and work I have done in collaboration with Hesburgh's specialist librarians, I have grown a new interest in NLP and qualitative text research, gaining knowledge and skills in techniques such as topic modeling and sentiment analysis. I am now able to continue expanding my knowledge of text mining and Natural Language Processing techniques, which allows me to continue introducing more qualitative styles of analysis into more traditional modes of empirical economic research.