More Than Just Cases II: Using Artificial Neural Networks to Map COVID-19 Social Media Sentiment

by John Hessler

Cartography has changed a great deal in the last twenty years and continues to evolve as new techniques from artificial intelligence, machine learning, and neural networks have begun to enhance the kinds of data that are possible to map. The definition of the object that counts as a map has also changed, from what was once a static representation of space, showing location and limited amounts of thematic information, to a dynamic temporal representation with multiple layers of complex data. Maps such as these allow cartographers and analysts to look deeply into natural and humanmade phenomena, which are rapidly evolving in space and time, and are often far from equilibrium.

Such is the case when cartographers approach the quickly-changing content of today's social media networks. Currently there are more than 500 million tweets written around the world per day and more than 800 million active users on networks like Instagram. These platforms generate, each hour of every day, a staggering amount of data that can tell us a great deal about what it is people are thinking, doing, and discussing around the world.

In the case of the current COVID-19 pandemic, caused by the SARS-CoV-2 virus, this data can be a critical source of information about how large groups of people are reacting to the virus and how the differences in their thinking about the disease vary in time and across large and small geographic regions.

New artificial intelligence methods, like Natural Language Processing (NLP) and neural networks, when combined with user geographic and location information are giving policy makers and epidemiologists unprecedented insight into how large portions of the population are reacting to the events now unfolding before us, in real-time.[1]

Sentiment analysis, which is also known as opinion mining, is a subfield of NLP. The basic purpose of sentiment analysis is to identify and extract opinions in a large corpus of texts. When these texts also contain geographic information, like that which can be extracted through the location services enabled by cellphone and computer users across the globe, cartographers and GIS practitioners can map the varying opinions of populations across wide swatches of geographic space.

NLP methods extract the opinions from text strings, like Tweets, by employing a concept called *word embedding*.[2] The computer treats words as vectors, whose similarity to other words, in a string or sentence, has been learned by an artificial neural network. The artificial neural network, which is made up of many layers of neurons has previously been trained with huge amounts of text allowing it to learn similarities and how each word is used in many different contexts.[3]

Each word is therefore represented in an artificial neural network by a real-valued vector that often has hundreds of dimensions. A word is converted into a row of numbers where each number is a dimension of the word's meaning and where semantically-similar words have similar valued vectors. What this means in practice is words like *father* and *man*, or *mother* and *woman*, would have more similar vector values than say, *truck* and *pig*.

Many sentiment-analysis algorithms have been developed over the last few years and an important one, called VADER, (Valence Aware Dictionary and sEniment Reasoner), employs a lexicon, or dictionary of sentiments, and a rule-based model for employing it across small strings of text.[4] The algorithm VADER is optimized for analyzing sentiments in short strings of words, such as those in social media posts like Twitter.[5] Several groups of researchers have started to employ these algorithms, including, the COVID-19 Infodemics Observatory, a project of the CoMuNe Lab, in Trento, Italy, to look at the variations and opinions of people across the globe, who are now facing the COVID-19 pandemic.[6]

VADER uses its learned dictionary to map emotional intensities written in the texts to sentiment scores. For example, words like *love*, *happy*, and *best*, all convey positive sentiment and are given a positive score. Lexicon and rule-based sentiment-analysis algorithms like VADER express this score by a polarity measure, which assigns a value in the range from −1 to +1, going from negative to positive. The algorithm is also smart enough to understand that strings like "did not like" express negative sentiments and emphasis notations such as capitalization ("GREAT") imply greater sentiment in one direction or another.

If we use the sentiment data generated by VADER with powerful mapping tools and the location information that geo-locates the source of a Tweet, we can begin to combine millions of these small and seemingly-independent expressions of opinion to get an idea of how different populations are reacting to the pandemic over time.

If for example, we take the 448.8 million geo-located Tweets that mention COVID-19 since the pandemic's beginning, we can generate a sentiment index for each country whose Tweets appear in significant numbers in the sample. A map of the current worldwide sentiment based on this algorithm can be seen in **Figure 1**. When looking at the map one can see that the United States has the most negative score compared with the rest of world.

Drilling down into the data for a few sample locations like the US, France, and Spain, we can express the sentiment score found in the geo-located Tweets as a time series for each of those geographic regions. The US graph [**Figure 2**] shows, since the pandemic began to the current moment, almost all the geo-located Tweets are negative.

France [**Figure 3**] and Spain [**Figure 4**], on the other hand, show a very different opinion profile with initial negative opinions turning mixed, and then in the case of Spain, almost completely positive, as the country's cases went lower during their extreme lockdown and later as they began to loosen restrictions on movement.

Mapping sentiments using NLP and neural networks is not perfect, and many difficulties remain with the technology including the ethical nature of using the data to generalize opinion over geographic regions. That being said however, the technology is presenting researchers with many questions about large scale phenomena, like the COVID-19 pandemic, making them contemplate the

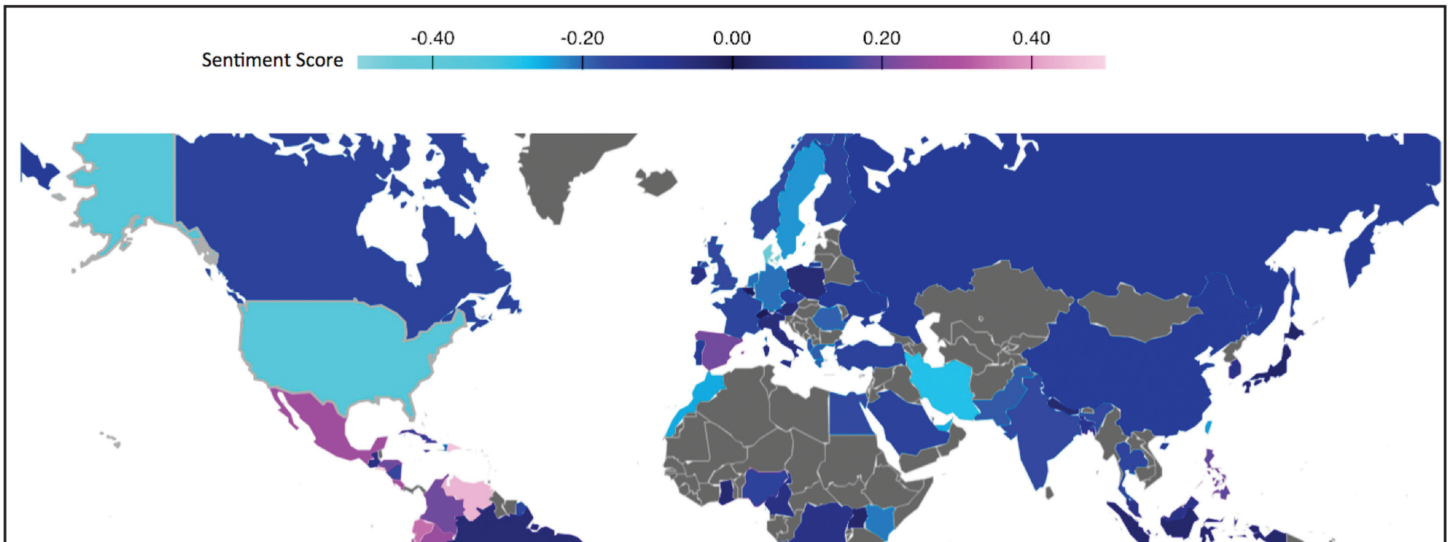

Figure 1: Partial World Map of Sentiment Using Lexicon Based Methods. Courtesy COVID-19 Infodemics Observatory.

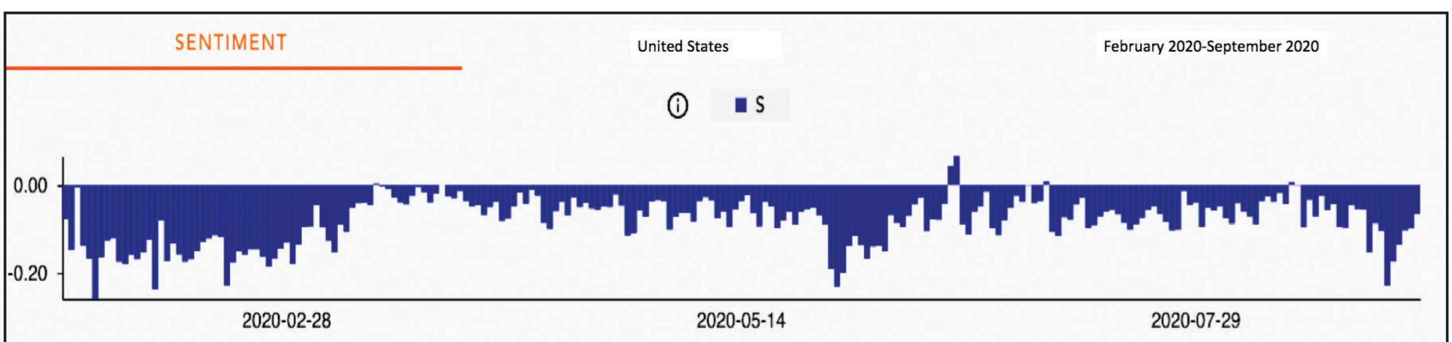

Figure 2: Time Series Sentiments for the United States. Made from COVID-19 Infodemics Observatory Data.

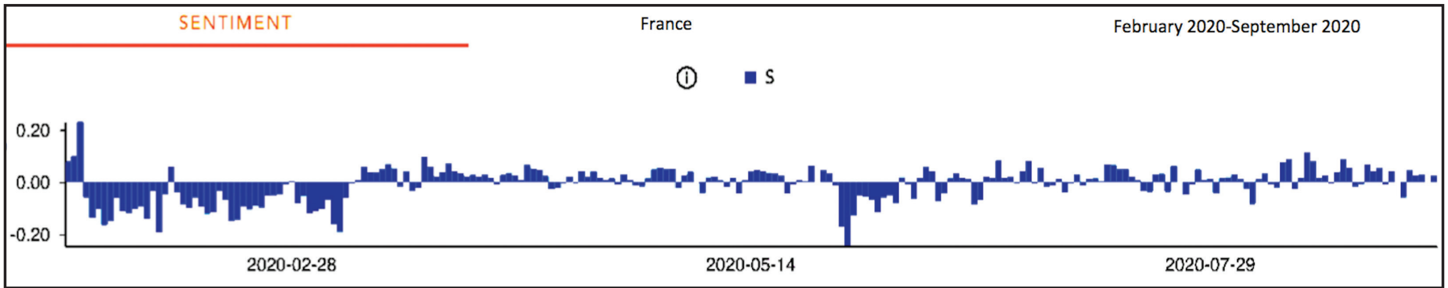

Figure 3: Time Series Sentiments for France. Made from COVID-19 Infodemics Observatory Data.

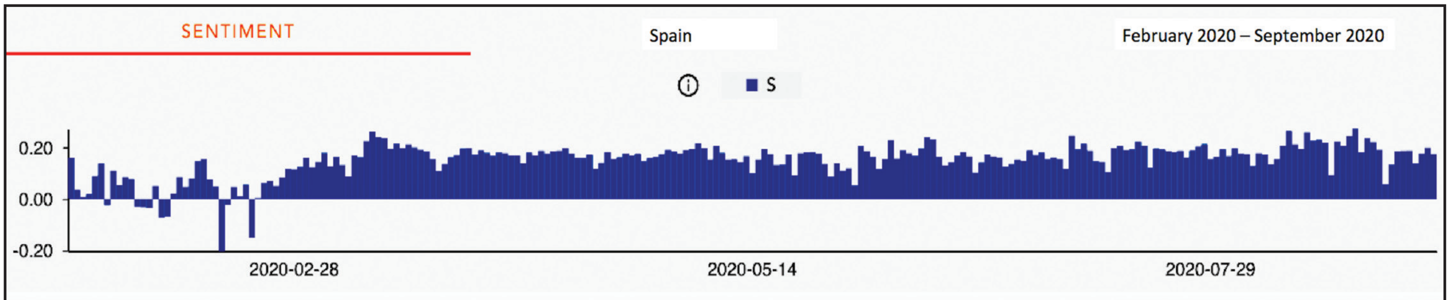

Figure 4: Time Series Sentiments for Spain. Made from COVID-19 Infodemics Observatory Data.

possible political, causal, and policy reasons for the varying differences in sentiment around the world.

Sentiment analysis is an example of the tools from NLP and artificial intelligence that will continue in the coming years to change increasingly what cartographers do, and what kinds of data they can bring to bear on spatial phenomenon. These technologies, combined with better data collection and machine learning, are opening up to mapmakers new ways of visualizing our world that only a few years ago would have been unimaginable.

—John Hessler is a Specialist in Computational Geography and Geographic Information Science at the Library of Congress. He is a Lecturer of Evolutionary Computation at the Johns Hopkins University and founder of the New Geometries Laboratory whose research uses geospatial data in conjunction with evolutionary computation to solve complex spatial analysis problems, like mapping COVID-19. This is his second article in a series on various mapping that supports understanding of the pandemic.

ENDNOTES

1 Jacob Eisenstein, Introduction to *Natural Language Processing* (Cambridge: MIT Press, 2019): 1–10.

2 Mike Thelwall, et al., "Sentiment Strength Detection in Short Informal Text," *Journal of the American Society for Information Science and Technology* 61, no. 12 (2010): 2544–58.

3 Apoorv Agarwal, et.al. "Sentiment Analysis of Twitter Data," *Computer Science at Columbia University*, accessed September 7, 2020, *http://www.cs.columbia.edu/~julia/papers/Agarwaletal11.pdf*; published In *Proceedings of the Workshop on Languages in Social Media* (LSM '11) (Association for Computational Linguistics, 2011): 30–38, *https://dl.acm.org/doi/10.5555/2021109.2021114*.

4 Maite Taboada and Julian Brooke, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics* 37, no. 2 (2011): 272–74.

5 Andranik Tumasjan, et al., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (Menlo Park, CA: The AAAI Press, 2010): 178–85.

6 The COVID-19 Infodemics Observatory is using lexicon-based NLP to analyze geo-located Tweets from around the world. It is their data pictured in this paper: *https://covid19obs.fbk.eu*.