Foundations of Data Science

Answer Template for CW2: Critical Evaluation of a Data Science Study

Which Study did you choose?

* [Predicting Disease Transmission from Geo-Tagged Micro-Blog Data](http://www.cs.rochester.edu/u/www/u/kautz/papers/Sadilek-Kautz-Silenzio_Predicting-Disease-Transmission-from-Geo-Tagged-Micro-Blog-Data_AAAI-2012.pdf)
* [Real-time tracking of self-reported symptoms to predict potential COVID-19](https://www.nature.com/articles/s41591-020-0916-2?fbclid=IwAR0tkHR2rBZ9vPGhoc8pm7ZreHZ0HdkVOzLGU2mxvYhgZ2n9DkWDZLRfys4)
* [Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study](https://jech.bmj.com/content/early/2020/11/10/jech-2019-213033?hwshib2=authn%3A1608649463%3A20201221%253Aa728626a-ac58-4878-ad87-4f3470ae783b%3A0%3A0%3A0%3ATJyrcbieqlV0%2Ftji4RhMwg%3D%3D)

The study I choose was to predict disease transmission from geo tagged micro blog data.

1. The following questions relate to the scientific paper of your chosen study. Note that the average length of a sentence is 15-20 words, the average length of a paragraph has between 3 to 6 sentences.
   1. [Basic] What is the scientific goal of the study? Write one to three sentences.

The scientific goal of the study was to be able to predict disease transmission by predicting the health of individuals from their geo-tagged micro-blog data (tweets on Twitter).

* 1. [Basic] What is the type of study (e.g., randomized trial, prediction model)? Include a brief explanation why it falls onto that category. Write one to three sentences  
       
     The type of study is a prediction model as a model was developed that was able to make predictions about individuals health state based on their location tagged tweets.
  2. [Advanced] What are the hypotheses of the study? Write one paragraph (3-6 sentences).

The hypothesis of the study was that by combining individual’s co-location and social ties with other individuals based on their Twitter posts leads to a higher performance of the prediction model in predicting if and when individuals will fall ill.

* 1. [Advanced] What is the methodology of the study and what is the results (how do the researchers go about testing the hypothesis, and what is the outcome? Write two or three paragraphs.

The methodology of the study involved the researchers using the Twitter search API to collect a sample of public tweets that originated from the New York City metropolitan area. The collection period was one month long which started on May 18, 2010 and by using a Python script, they searched Twitter for all recent tweets within 100 kilometers of the NYC center. Overall, the researchers collected nearly 16 million tweets by more than 630 thousand unique users and they specifically focused only on those specific accounts that posted more than 100 tweets with their location tagged during the one-month data collection period. Their dataset consisted of 6,237 such individuals.

So, before the researchers could train and evaluate a predictive model of personal health, they needed to identify individuals who were ill and estimate the time as to when they became contagious. So, they focused on symptoms that were self-reported and on complaints that appeared on Twitter status updates. The researchers mentioned that they “learned a linear support vector machine SVM binary classifier which we used to detect tweets of individuals who were sick” (quote from study). The researchers needed to obtain a sufficient amount of labeled training data to learn their SVM model, so they trained two helper SVM’s on a dataset of around 5,128 tweets and then they used these helper SVM’s again to label a set of around 1.6 million tweets that “were health related but contained some noise” (quote from study). The researchers would only consider two individuals being co-located if they were within a 100x100 meter cell within a certain time window. The researchers found an “exponential relationship between the probability of individuals getting ill and their co-location” (quote from study) so they wanted to model the joint influence of co-location and social ties so that they could predict the health state of any individual on a certain day. The researchers also mention that they “learned a CRF model that would model the health of an individual over a number of days and observe each day, day of the week, history of sick friends in the past, recent co-location with other individuals who were sick and also the number of such individuals that were encountered” (quote from study).

The results of the study showed that when “leveraging the effect of relationships or co-locations individually the CRF model performed inconsistently as further predictions into the future were made by the researchers” (quote from study). On the other hand, when the researchers considered friendships and co-locations jointly the performance of the CRF model was more stabilized and improved. Thus, the results of this study confirmed the hypothesis mentioned earlier.

* 1. [Intermediate] What are the statistical methods used in the study and how are they applied to the data? Write one or two paragraphs.   
       
     One of the statistical methods used in this study was hypothesis testing. This was used as the researchers wanted to test whether the model would perform well if both the friendships and co-locations were used or if they were individually used. This is applied to the data as graphs were used to show precision and recall which analyzed the performance of the prediction model to determine the accuracy of the predictions made.
  2. [Basic] What are the stated conclusions of the study? Write one paragraph.

The researchers mentioned that the research conducted was “the first to take on prediction of the spread of infectious diseases throughout a real-world population with fine granularity” (quote from study). The researchers mostly focused on symptoms that were self-reported and appeared on an individual’s Twitter status. They showed that although these messages were rare, they were still able to identify them with high precision and high recall through their model. “The key contribution of this work is a scalable probabilistic model that demonstrates that the health of a person can be accurately inferred from their location and social interactions observed via social media” (quote from study). The researchers managed to show that their model could predict the future health states of individuals with a consistent high accuracy more than a week into the future. They also mentioned how an early identification of infected individuals was crucial in being able to prevent and contain disease outbreaks. In conclusion, the study demonstrated that finding individuals along with others that may have contracted a disease could be done effectively and quickly based on their Twitter posts.

* 1. [Advanced] Provide a critical discussion of the paper. How important or impactful is it? Are there any obvious errors or potential flaws? Write one or two paragraphs.

The study was very import in my opinion the researchers were able to develop a successful prediction model that was able to predict possible disease outbreaks in the future with high accuracy. The study was also very impactful in my opinion because by using the prediction model the researchers can predict possible disease outbreaks just by looking at the health of individuals within a geographical area. This would then allow action to be taken which would help to contain a pandemic disease/virus or even prevent it from occurring in the first place based on the prediction of the model.

One of the limitations mentioned in the study was that the observations made were limited by the commonness of public tweets in which the users talked about their health and “by our ability to identify them in the flood of other types of messages” (quote from study). So, this led to an underestimation being made of the number of individuals that were infected.

* 1. [Intermediate] Are there any ethical implications of the study? If none, please state so; if yes, how well do the authors relate to these implications? Write one paragraph.   
       
     There are few ethical implications in this study. One of the major implications is of informed consent. As the researchers use Twitter posts of real individuals to train their prediction model without the individuals being aware of as to what their data is being used for. From examining the study paper, I found that the researchers do not relate to any ethical implications at all as they do not raise any ethical issues and ways that they look to tackle them.

1. The following questions relate to the media report of your chosen study:
   1. [Basic] Provide a brief summary of the report. Write up to three sentences.

The report article mentions how the model was able to track diseases as it spreads in a city. It also mentions how the AI algorithm was correct in its prediction around 90% of the time and made predictions 8 days in advance. Lastly, the article mentions how the AI algorithm system was able to monitor the health state of individuals and their friends.

* 1. [Advanced] How accurately did the report summarize the study? Write one or two paragraphs.

Overall, the article does a pretty good job in summarizing what the main aspects of the study were. For example, it mentions the figures correctly like the number of tweets that were collected along with their tagged locations analyzed. It also mentions accurately as to what methodology was used by the researchers such as training an AI prediction model and what this meant in the long term i.e., being able to improve their understanding of the emergence of pandemics globally based on the day-to-day interactions of individuals.

However, that being said the article mentions some points that are not found in the study. One such example of a mention is that in the article it states the following: “In unpublished findings described to New Scientist during an interview at the Conference on Artificial Intelligence in Toronto Canada, the team also revealed that people who go to the gym regularly are moderately less likely to get sick. People with low socio-economic status, on the other hand, are much more likely to become ill.” As this statement was from unpublished findings and was not mentioned in the study itself, this suggests the efficacy of this statement cannot verified by the study. Although, the use of heat maps to show the spread of flu through a particular area in New York was not mentioned in the study, the study still mentioned using maps to represent friendships and co-locations amongst the users. However, I think that as the ultimate objective of the prediction model was to predict upcoming epidemics in a specific area e.g., New York based on the tweets of individuals which the article is able to convey in a better and simpler manner through the use of heat maps.