**Final Project #2 - Tiffany Li**

**Project Problem and Hypothesis**

* **What's the project about? What problem are you solving?**

Privacy policies are legal obligations for most websites and provide important information on user data privacy. However, policies are often difficult to read, too long, or incomplete (lacking information that is legally necessary or industry standard). My hypothesis is that factors like length of policy and readability index of policy will be positively correlated with completeness of privacy policy. So, you would be able to roughly predict the completeness of a privacy policy by quickly skimming the policy for length and readability.

* **Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?**

I think linear regression would show the relation between the variables.

* **What kind of impact do you think it could have?**
* **What do you think will have the most impact in predicting the value you are interested in solving for?**

This could have interesting implications for privacy policy design. Findings may be relevant to tech companies, privacy advocates, and consumers.

**Datasets**

* **Description of data set available, at the field level. (see table)**
* **If from an API, include a sample return. (this is usually included in API documentation!) (if doing this in markdown, use the JavaScript code tag)**

Pull policies from top websites by Alexa rating or other metric. Convert to machine readable text to analyze length and readability rating. I would use the same methods found in the papers I cite below, some of which converted policies to machine readable text, analyzed length, and analyzed readability rating.

For “completeness,” I would conduct a subjective analysis by reading through each policy and looking for presence of specific factors, based on privacy frameworks.

**Domain knowledge**

* **What experience do you already have around this area?**

I am a privacy law expert. As a technology lawyer, I’ve worked with startups and large tech companies, including Amazon, Ask.com, and Wikimedia (the non-profit that runs Wikipedia). I have numerous privacy certifications and honors, and I am a Women Leading Privacy Advisory Board Member for the International Association of Privacy Professionals. I am also an Affiliate Scholar for Princeton’s Center for Information Technology Policy, where my research currently focuses on privacy and intellectual property as applied to artificial intelligence.

* **Does it relate or help inform the project in any way?**

My domain knowledge is highly relevant to this project. In fact, this project came out of my experience working in privacy law. I have personally written dozens of privacy policies and reviewed even more. Some of the work involved in this project will be very specific to current privacy practices and will require knowledge of privacy law and recent research in privacy.

* **What other research efforts exist?** 
  + **Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.**
  + **This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."**
  + **Include a benchmark, how other models have performed, even if you are unsure what the metric means.**

I conducted a pre-emption search on this topic, and I did not find anything recent that comprehensively looks at the problem in the way I propose. Here are a few studies that helped inform my proposal:

1. Elisa Costante, Yuanhao Sun, Milan Petkovi´c, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In Proceedings of the ACM Workshop on Privacy in the Electronic Society. <http://repository.tue.nl/748780>
2. Reidenberg, Joel R., Breaux, T. D., Cranor, L. F., French, B., Grannis, A., Graves, J. T., Liu, F., McDonald, A. M., Norton, T. B., Ramanath, R., Russell, R. C., Sadeh, N., and Schaub, F. [Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding.](http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2418297) Berkeley Technology Law Journal, 30(1), May 2015, 39-88.
3. McDonald, A. and Cranor, L. The Cost of Reading Privacy Policies. I/S: A Journal of Law and Policy for the Information Society. 2008 Privacy Year in Review issue.
4. Joseph D. Brown, Mohammad A. Ghani, Muyeedul Hoque, Umair A. Rehman, An Analysis of Web Privacy Policies Across Industries, Worcester Polytechnic Institute. <https://web.wpi.edu/Pubs/E-project/Available/E-project.../IQP_Final_Report.pdf>
5. Katy Steinmetz, “These Companies Have the Best (And Worst) Privacy Policies.” Time Magazine. August 2015. <http://time.com/3986016/google-facebook-twitter-privacy-policies/>

Essentially, there have been comparative empirical studies on privacy policy length, privacy policy readability, and privacy policy completeness (presence of necessary elements according to various privacy laws). I have not found any studies that have looked at policy length, readability, and presence of various elements as factors that may be related in ways that would allow conclusions to be drawn.

**Project Concerns**

* **What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help)**

I’m not sure how to measure “completeness” (presence of certain elements). I could do: T/F complete or not complete, # of elements present, % complete, or look at each element separately. The other remaining variables are discrete – number of words, readability score.

* **What are the assumptions and caveats to the problem?** 
  + **What data do you not have access to but wish you had?**
  + **What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample. (say, predicting who would win an election)**

One caveat is that I will be ignoring policies that are broken up into multiple pages, because it is difficult to compare those with single-page policies.

Ideally, I would like independent verification of completeness (presence of certain elements) in addition to my own analysis, but that would take too much time for this project.

Already implied: Small sample, may not be representative of entire industry or of all privacy policies. Clearly subjective bias in selection of policies, selection of elements (though mitigated b/c I will be basing those on elements noted in laws).

* **What are the risks to the project?** 
  + **What's the cost of your model being wrong? (What's the benefit of your model being right?)**
  + **Is any of the data incorrect? Could it be incorrect?**

It’s possible that there will be no meaningful conclusions drawn from this study and there might be no relationship between length of policy and readability vs. the other variable/s. It is also possible that selection bias would make any results useless in terms of generalizing to larger conclusions.

Data could be incorrect, especially given subjective bias + lack of independent confirmation.

**Outcomes**

* **What do you expect the output to look like?**
* **What does your target audience expect the output to look like?**

I expect there to be a relation between length of policy and readability, and between those two variables and the variable/s of “completeness.” I would expect some people in my target audience would assume that longer policies = more complete, but I’m guessing the opposite.

* **What gain do you expect from your most important feature on its own?**
* **How complicated does your model have to be?**
* **How successful does your project have to be in order to be considered a "success"?**
* **What will you do if the project is a bust (this happens! but it shouldn't here)?**

I think any sort of comparison will be a success, even if the results are simply that there is no relation between any of these variables and you cannot predict anything based on an of the variables. Even if the project is a bust, it will be useful to have the data.