Application for Predicting Targeted Content Online

(Technical Paper)

Attitudes Towards the Valuation of Data Privacy on Social Media

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia | Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Daniel Le

Fall, 2020

Department of Computer Science

Signed: Daniel Le Date: 12/11/2020

Approved: _____ Date: _____

S. Travis Elliott, Department of Engineering and Society

Approved: Aaron Bloomfield Date: 12/11/2020

Aaron Bloomfield, Department of Computer Science

Introduction

In the age of accessible information, social media reigns supreme as one of the most dominant mediums of communication, eclipsing even television and newspaper (Broesma and Graham, 2012). Thoughts and reports can be relayed across the planet in a matter of milliseconds. Entertainment such as videos, music streaming, and educational content are available at the touch of a button. Even commerce now trends towards online transactions rather than brick-and-mortar shopping, with 77% of consumers preferring the former (Maat, 2018). Despite the cost of most major platforms being free, the companies that own them are among the most valuable in the world. Facebook holds a firm position as the fifth most valuable entity on the planet (Corrigan et. al., 2018). Most revenue is therefore generated from the users, in the form of data mining and target advertising. But is this implicit exchange of virtual media for personal information warranted, or should both involved parties place more consideration into their choices?

The potential technical project is oriented towards giving users more insight about how their online behaviors impact the advertisements and targeted content they receive, and how it affects their activity as a money spender. I will implement an application that aggregates their interactions across platforms, as well as their recent purchases in their daily lives. Users will be able to manipulate these factors in an isolated setting. They will be able to see the correlation between the two datasets, and gain insight as to how companies may view them as the product, not the customer.

My STS research topic aims to analyze the importance of data privacy to social media users, as well as the reasons as to why these companies have access to so much of that data. The Social Construction of Technology (SCOT) framework will be crucial in interpreting users'

attitudes towards relinquishing privacy rights. What factors do they emphasize when protecting their privacy, and what are the possible responses to social media companies that breach the relationship of trust? Are individuals inclined to stand up for their rights, or are there worthwhile tradeoffs that make it permissible to blur the limits of privacy? I will also examine the generational trends that incline social media users to place less value on their positions as vital stakeholders in the industry.

The prevailing goal of this thesis is to promote more vigilance from social media users regarding the impacts of their actions. As consumers, we provide social media companies our data, and in return we receive higher quality services from them. That exchange can be expected, but data collection has increasingly been abused, with that same data being used to cultivate propaganda, bias, and misinformation, like in the Cambridge Analytica-Facebook scandal (Hu, 2020). Whether or not they accept having their data exploited is subjective, but the thesis should at least clarify how much impact they can have as an individual.

Technical Topic

Introduction

Social media users may simply perceive their likes, comments, and views as ways to interact with others, but often these seemingly minor interactions have deeper meanings for the companies providing the platform. One's trend of behaviors shapes their virtual footprint and reveals their interests as a potential consumer of goods and information. While there are reasonable benefits to this, such as suggesting relevant products to buy and making browsing through content more convenient, there also exists the potential for abuse. If users wish to experience different points of view or simply avoid biased content while still being able to use the platform to its potential, how can they know where and how their statistics are being processed?

This technical project shall allow users more insight into their digital footprint. It will utilize datasets of several criteria: likes, comments, search terms, link clicks, and shopping cart/purchase histories (should we have access to that). With proper formatting and manipulation, the data could be fitted to a model that predicts the most likely categories and companies that the user behavior appeals most to. Combined with values about follows and views, users can visualize the impact of their actions and the efficacy of targeted content.

While currently small in its scale, the project could have implications for up to half of the world's population, considering there are 2.38 billion active monthly users of Facebook alone (Appel et. al., 2019). Some companies are more transparent than others when it comes to storing personal data, but in most cases, it is difficult, if not impossible, to reveal how someone is marketed to advertisers. Being able to understand the direct consequences of one's clicks can help to reduce the negative effects of biased content.

Methods

The project draws upon topics covered in the CS 4774 machine learning course. The nature of the intended dataset suggests an implementation of a classification algorithm. It will be written using Python and saved into a Jupyter notebook. Given a user's social media behavior, the model will predict the market that they most closely align to. For example, if there is a user that frequently interacts with tech pages, searches for news about the latest cell phones, and has recently bought several charging adapters, then the machine learning model may likely suggest the user may be targeted by companies selling phone cases.

In order to get the most accurate predictions, the model must be trained with previously determined data. This could be sourced from previously collected sets online, from reputable databases such as *Kaggle.com.* Additionally, inputs could come directly from asking around for willing participants. Both these sources must consist of information on both user behavior as well as the targeted content they consequently receive. An important consideration is that the input values must all be numerical. Amounts such as likes and followers are easily gathered, but categorical data like keywords must be transformed into integers using one-hot encoding. Once the arithmetic is applied to the numbers, the resultant value can be matched to a categorical result, which would be one of the many predetermined "customer alignment" labels.

The collection of data will likely be trained in random batches to maximize the variability across a relatively limited scope. Random training groups allows for more unique combinations to improve the machine learning model's accuracy without requiring wholly new datasets for each iteration.

Being a small-scale project that relates to a field involving thousands of advertisers and billions of dollars of research, it may not be possible to perfectly apply the classification model

to every given input. With the timeline limited to just the Spring 2021 semester, the goal is to produce a working application, with expected room for improvement regarding accuracy and simulation speed. Ideally, a proof of concept shall be ready come March, and a minimum viable product with basic UI elements by May.

STS Topic

Introduction

While the technical project is designed to alleviate the concerns regarding targeted content and digital footprints, the STS research project shall analyze the social factors as to why the problem has arisen in the first place. The analysis will be twofold – it is necessary to look at what motivates companies to benefit from users' personal data, as well as the attitudes from the users themselves that enable it. How important are privacy rights to the average consumer? Why are they disposed towards waiving these rights so freely?

STS Framework

The social trends that cause individuals to value their privacy either more or less is best covered by the Social Construction of Technology (SCOT) framework. An important distinction that SCOT makes is that human behaviors shape the progress of technology, not the other way around (Cengage, 2020). The technology in question cannot be analyzed by itself; its performance is dependent on the people that utilize it. It could be inferred that the issues of privacy are not *caused* by social media, rather social media is just a catalyst that exposes the flaws among the userbase. SCOT provides formal criteria to which social media and privacy rights can be evaluated, rather than blindly trying to extract relevant points of contention from a vast source of information.

The interpretive flexibility principle of SCOT states that the same technology can represent different meanings for different people. With quite literally almost half of the world being involved with some form of social media, there are expectedly a plethora of different attitudes towards how invasive the platforms can be. Just some of the relevant social groups include avid content creators, infrequent status updaters, the app developers, and the companies

that wish to send out their advertisements. Each has a different goal in mind, with varying levels of concern towards privacy and fair usage. The younger demographic that values easy entertainment over most else would likely be less protective of their personal data, so long as they can receive relevant updates about their favorite online sources. Older adults who grew up before the digital era may prioritize their individualism more, and aim to leave as little presence on social media as possible (Hoofnagle, 2010). And finally, the corporations that provide the goods and services might claim that they respect user rights upfront, but they still have to make money to stay in business. They might want as much access as they can into users' lives, to better curate content that keeps them coming back, and to gain additional stakeholders in businesses that utilize the social media platform to promote their own businesses. This inclination can be seen before even using any of these services, as terms and conditions agreements are intentionally more complex than most classic literature in terms of readability (Luger et. al., 2013), so that users may unknowingly agree to legal loopholes through which their data can be exploited. The design flexibility subcategory can be applied to the current balance between data permissions and usability. Are the current market offerings the optimal compromise between minimally invasive practices and a satisfactory user experience? The STS research paper shall evaluate the design choices of certain apps, and how they could be improved to respect the average user more.

The completeness of these designs will impact the interpretation of SCOT's closure principle. As each problem concerning the proliferation of user data is gradually resolved, the need for improvement declines accordingly. Users of social media can always identify further conflicts regarding the platform and its practices, which lead back to the evaluations under interpretive flexibility. To achieve closure requires a shift in the attitudes from at least one of the

sides of the debate about privacy. One must concede and allow the social norm to adjust to that new standard. Are better anonymization algorithms, where data is still processed, but without associating it to a name, the solution (Bakken, et. al., 2004)? Or perhaps limiting access in the first place through laws that "mandate that entities that collect or process data… must delete any data relating to an individual…upon request" (Victor, 2013)? The research paper will consider further approaches to this goal, and how much say the average social media user has in the process.

Plan and Methods

Once again, with so many stakeholders in this topic, extensive elicitation of perspectives is necessary. First and foremost, qualitative reasoning is important to understanding the status quo. Data privacy is inherently a moral debate, and opinions may be just as important as numbers. One source of information is simply just asking peers how they feel about how their data being used online beyond their immediate knowledge. This can be completed through surveys and interviews. In addition to the questions that I personally ask, I will search for interviews available online. There are people with similar concerns that have access to a broader audience, both when it comes to more users of social media and representatives of the companies that use personal data for profit.

To support the claims from those firsthand accounts, more information will be sourced from publicly available values about ad revenue and usage statistics. This is relevant to determining whether the concerns highlighted in the interviews are founded, and if so, how deep does the problem of data as a commodity go. It will provide an indication to potential targets for a solution.

The testimonies gathered from stakeholders will serve as examples of interpretive flexibility. Opinions expressed in each account will provide an insight into a subset of *what* attitudes exist towards data privacy, while surveys and polls can demonstrate *to what extent* they are felt. For example, do a majority of adolescents feel indifferent towards disclosing their data in return for more personalized feeds? Together with extrapolations from the usage statistics, the most profitable aspects of user data can be identified. If these criteria align with what users feel more protective of, it would suggest changes are needed to the design model in order to progress towards closure as defined by SCOT. Conversely, if the more profitable factors are not what users identify as problematic, does it suggest closure is more attainable than expected? Or perhaps should more concern be warranted by the users?

Conclusion

The research conducted during both the technical and STS projects will hopefully result in a better understanding society's perception towards data privacy on social media. Should people have reason to be protective of their personal information? Is relinquishing this information worthwhile as a tradeoff for more effective communication methods? What responsibilities do businesses have when it comes to making a profit yet respecting the rights of their consumers? Keeping these questions in mind, and combining them with data collected through research and elicitation, perhaps we can explore better options to improve the relationship between individuals and social media companies.

References

Appel, G., Grewal, L., Hadi, R. *et al.* The future of social media in marketing. *J. of the Acad. Mark. Sci.* 48**,** 79–95 (2020). https://doi.org/10.1007/s11747-019-00695-1

D. E. Bakken, R. Rarameswaran, D. M. Blough, A. A. Franz and T. J. Palmer, "Data obfuscation: anonymity and desensitization of usable data sets," in *IEEE Security & Privacy*, vol. 2, no. 6, pp. 34-41, Nov.-Dec. 2004, doi: 10.1109/MSP.2004.97.

Compaine, B. M., & Vogelsang, I. (2000). *The Internet upheaval: Raising questions, seeking answers in communications policy*. Cambridge, MA: MIT Press.

Corrigan, J. R., Alhabash, S., Rousu, M., & Cash, S. B. (2018, December 19). How much is social media worth? Estimating the value of Facebook by paying users to stop using it. Retrieved October 18, 2020, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207101

E. U., Luger, T. O., Nottingham, S. U., Stuart Moran The University of Nottingham, T. U., & Tom Rodden The University of Nottingham, U. U. (2013, April 01). Consent for all: Revealing the hidden complexity of terms and conditions. Retrieved from https://dl.acm.org/doi/abs/10.1145/2470654.2481371

Hoofnagle, Chris Jay and King, Jennifer and Li, Su and Turow, Joseph, How Different are Young Adults from Older Adults When it Comes to Information Privacy Attitudes and Policies? (April 14, 2010). Available at SSRN: https://ssrn.com/abstract=1589864 or http://dx.doi.org/10.2139/ssrn.1589864

Hu, M. (2020). Cambridge Analytica's black box. *Big Data & Society*. https://doi.org/10.1177/2053951720938091

Kees Maat, R. K. (n.d.). Accessibility or Innovation? Store Shopping Trips versus Online Shopping - Kees Maat, Rob Konings, 2018. Retrieved October 18, 2020, from https://journals.sagepub.com/doi/full/10.1177/0361198118794044

Marcel Broersma & Todd Graham (2012) SOCIAL MEDIA AS BEAT, Journalism Practice, 6:3, 403-419, DOI: 10.1080/17512786.2012.663626

"Social Construction of Technology ." Encyclopedia of Science, Technology, and Ethics. . Retrieved October 16, 2020 from Encyclopedia.com: https://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/social-construction-technology

Victor, J. M. (2013, November). The EU General Data Protection Regulation: Toward a Property Regime for Protecting Data Privacy. Retrieved October 18, 2020, from https://www.yalelawjournal.org/comment/the-eu-general-data-protection-regulation-toward-a-property-regime-for-protecting-data-privacy