

University of Oxford



Master's Dissertation

To Become a Mockingbird:
Providing Twitter Users with Tools to Control Algorithmic Perceptions

Author: Adam Hare
Advisor: Professor Max Van Kleek

*Submitted in partial fulfillment
of the requirements for the degree of
Master's of Science*

Department of Computer Science
University of Oxford

3 September 2019

Abstract

As algorithms are increasingly used to make potentially life-altering decisions the importance of ensuring they meet certain standards of fairness, accountability, and transparency is becoming paramount. Complex models that are poorly understood are susceptible to bias, relying on protected or irrelevant features, and adversarial examples. This concern has spawned interest in explainable and interpretable algorithms which allow humans to understand algorithmic decisions. Running parallel to this issue is an increasing reliance on algorithmically generated profiles to target advertisements online. Often these profiles are built on sensitive characteristics, without user knowledge or means of recourse. They can influence user perceptions of self, alter the content users are exposed to, and reveal to other parties characteristics which users have chosen not to share. This project aims to increase user autonomy and privacy online by leveraging techniques in explainable machine learning.

This work begins by creating a prototype tool called Mockingbird that builds more than twenty algorithmically generated profiles based on Twitter data. These profiles are created using various techniques, including lexicons, neural networks, and Gaussian Processes. Mockingbird offers explanations for most of these profiles, primarily using a post-hoc explanatory tool. Finally, using synonym suggestions and techniques used to generate adversarial examples, Mockingbird gives users suggestions on how to alter their tweets in order to change their algorithmic profiles. Mockingbird is novel in combining profiles, explanations, and tools to change algorithmic profiles.

The system was tested through several ($n = 6$) lab experiments to gauge user response to explanations and willingness to alter their data. It appears that users are generally not concerned about their algorithmic profiles when used for advertising, but would be much more concerned in higher stakes applications. Users were also unwilling to change their existing data to control algorithmically generated profiles, viewing the necessary changes as too disruptive to how they use social media. This work points to avenues of further research focusing explicitly on editing social media data to influence high stakes decisions such as job, visa, and loan applications.

Acknowledgments

There are several people I would like to thank for their help and support throughout this project. First, thank you to my supervisor, Max Van Kleek, for all of your help, especially in selecting the specific research question, helping me to set up the Mockingbird tool, and your guidance with the ethics approval and interviews. Thank you to my family for your unceasing support despite the distance and for having the patience to listen to my ideas, even though only a few of them panned out. Thank you to my friends at Oxford, especially Lisa for your companionship and advice, Sheryl for your kindness and support, and James for your friendship and distractions from the world of computer science. Thank you to all of my friends from the States - Luke, Jacob, Courtney, Cas, Ptom, Amol, Katie, and the rest. You've again proven that distance is no barrier to close friendships and I eagerly await seeing you all again.

This paper represents my own work in accordance with University regulations.



Adam Hare

Contents

Abstract	1
Acknowledgments	2
1 Introduction and Motivation	5
1.1 Motivation	5
1.2 Contributions	7
1.3 Research Questions	8
2 Background	11
2.1 Explainable & Interpretable AI/ML	11
2.1.1 Definitions	11
2.1.2 Problems & Open Questions	13
2.1.3 Explanations Through Approximation	15
2.2 Building Profiles from Social Media	16
2.2.1 Use of Social Media Data	17
2.2.2 Types of Profiles	19
2.2.3 Classification From Profile Data (M3)	24
2.2.4 User Response	25
2.3 Privacy Enhancing Technology	26
2.4 Self Curation - A Sociological Overview	31
3 Methodology - Prototype	35
3.1 Prototype Design	35
3.1.1 Generating Algorithmic Profiles	36
3.1.2 Explanatory Tools	41
3.1.3 Editing Tools	43
3.2 Prototype Implementation	44

3.2.1	Training Neural Networks	44
3.3	Income	47
3.3.1	Analysis of Editing Tools	47
4	Methodology - Experiment	53
4.1	Experimental Design	53
4.1.1	Recruitment	54
4.1.2	Special Category Data	54
4.1.3	Experiment Overview	55
4.2	User Interface	55
5	Results	63
5.1	Participant Information	63
5.2	Thematic Overview	64
5.3	Study Results	65
5.3.1	Expectations	65
5.3.2	Interest in Profiles	67
5.3.3	Explanations	72
5.3.4	Willingness to Change	73
6	Conclusion	76
6.1	Limitations	76
6.2	Conclusions	77
6.2.1	Addressing Research Questions	77
6.2.2	Implications for Future Work	79
6.3	Closing Remarks	81
	References	82

Chapter 1

Introduction and Motivation

1.1 Motivation

As society integrates technology into many varied aspects of daily life, people’s lives are frequently shaped by decisions made by algorithms. In some situations of high importance, such as loan applications and job recruitment, these algorithmic decisions may be life-altering. Recent legislation in the European Union has begun to address such issues, with Article 22 of the General Data Protection Legislation (GDPR) forbidding decisions based solely on automated profiling in cases where the decision “produces legal effects [...] or similarly significantly affects” an individual. [European Union, Parliament and Council, 2016]¹ Even less obviously, significant usage of these profiles may have wide-reaching effects: algorithms which target advertisements or curate the content of a social media newsfeed can alter user’s self-perception and mood both directly and indirectly [Kramer et al., 2014] in an attempt to influence their future behavior. These effects can be achieved without the awareness or consent of users.

Individual users are not the only ones who struggle to comprehend these systems. For many machine learning techniques, not even the programmers who developed the model are able to make sense of its decisions. The problem is compounded when non-specialists such as regulators and downstream users require an understanding of the model [Veale et al., 2018]. This opacity can make debugging difficult and increases the likelihood that a model with a serious flaw will be put into use before the issue is discovered. Especially in high stakes applications, models must be correct for the “right reasons,” as established by their creators and regulators [Hendricks et al., 2018]. If not handled properly, deep learning algorithms can suffer from indirect discrimination by learning

¹In the Preamble point 71, GDPR explicitly mentions loan applications and automated recruiting as protected cases.

proxies for protected characteristics to produce discriminatory decisions [Zliobaite, 2015].² Even less obviously salient applications, such as automated captioning technology, can perpetrate harmful biases from the training set [Hendricks et al., 2018, Garcia, 2016].

These issues have resulted in calls for new systems to conform to ethical design principles such as fairness, accountability, and transparency [Mittelstadt et al., 2016, Lepri et al., 2018]. Related to transparency is the concept of scrutability.³ A model or system is *scrutable* if it allows for some part of the decision-making process to be parsed by humans in a way that results in a reasonably useful understanding of how the decision is made. Two popular examples of this are the connected fields of explainable and interpretable machine learning (ML) and Artificial Intelligence (AI). Explainable models provide a human-understandable rationale for their decisions and interpretable models are such that the exact process they used to make a decision can be traced by a human (see Section 2.1.1). In spite of legislative interest, such as GDPR mandating that any automated profiling allow the subject “to obtain an explanation of the decision reached” [European Union, Parliament and Council, 2016], few scrutable systems have been widely adopted. Even when scrutable systems for high stakes applications are available, older inscrutable systems are the norm.⁴ Despite their appeal, there are several open questions regarding the usefulness and practical application of scrutable systems.

Interpretable and explainable models are not without their shortcomings. They may have only a limited ability to give users actionable steps to change the way they are perceived by algorithms. Research has found that many users respond more positively to explanations that are “actionable” [Binns et al., 2018]. Other research has attempted to design tools to provide individuals with “actionable recourse” in order to change their classification by an algorithm [Ustun et al., 2019].⁵ Work on actionable recourse has focused on creating new classifiers designed with this ability in mind, but this may not be a reasonable expectation for algorithms that are not as obviously life-altering or where the details of the classification need to remain secret. Additionally, while scrutable models can improve future applications, they are little help to users subject to algorithms today.

²An example would be using an applicant’s name or neighborhood as a proxy for race. Indirect discrimination can be difficult to combat in general, as it may be quite subtle, but using black box algorithms only exacerbates the problem.

³This is also referred to as “comprehensibility” in some papers, see [Rudin, 2018].

⁴Consider the criminal recidivism problem as addressed by COMPAS [Brennan et al., 2009], a black box algorithm in use in the United States, and CORELS [Angelino et al., 2017], an interpretable model that achieves comparable accuracy. COMPAS is used widely even though interpretable models exist and the application is clearly of high importance. See [Rudin, 2018] for a more detailed comparison.

⁵Although this work is limited to simple linear classifiers, in its extensions it suggests that non-linear classifiers could be examined with tools such as LIME [Ribeiro et al., 2016].

1.2 Contributions

This research intends to contribute a novel autonomy-enhancing tool⁶ which empowers users to control how they are perceived online without relying on changes from industry. Specifically this project focuses on helping users realize and alter algorithmic classifications based on their social media data. These data influences users' newsfeeds as well as the ads they are shown; both of these have been shown to impact users' mental state [Kramer et al., 2014, Park and Grow, 2008]. Additionally, certain sensitive characteristics such as religion and sexual orientation [Kosinski et al., 2013a], if profiled, could be revealed to others through targeted ads or suggestions. Similar approaches to the ones in this research could be used to help users combat negative profiling for job applications and dating sites. Such tools are particularly desirable as many people do not realize what these algorithms are profiling, let alone how they work or how accurate they are. Even if people are aware and seek information on their profiles, explanations offered by social media providers can be vague and misleading, and often offer no recourse [Andreou et al., 2018] (see Section 2.2.1).

This work aims to build upon existing research by combining current profiling and explanatory tools, measuring how users respond to them, and observing changes users make to control their algorithmic profiles. Research into the effects of profiling [Ur et al., 2012, Gou et al., 2014] and explanations [Stumpf et al., 2016, Andreou et al., 2018] has generated many insights into these topics individually but there has been little study regarding their combined effect. There is a similar dearth of work on the sorts of compromises users are willing or unwilling to make to change algorithmic perceptions. While automated privacy enhancing systems can be useful in extreme cases, they are not viable for every day users because they lack controls necessary for the realization of nuanced and individualized representations of self. Providing users with autonomy-enhancing tools to understand and control their own profiles respects both an individual's need to curate their online identity and desire to control algorithmic profiles. Although individuals may choose to control their social media in ways that do not optimally obscure algorithmic perceptions, they have been both informed and empowered to make this decision in new ways. Additionally, this approach allows users who are not domain experts to interact with current models and see results immediately, without waiting on changes from industry and without requiring formal training or a technical background.

To accomplish this task, I create a prototype application called Mockingbird which combines a wide array of algorithmically generated profiles based on Twitter data. This new and novel

⁶Autonomy-enhancing tools, a superset of privacy-enhancing tools, give users more control over the ways in which their lives are impacted by technology. These tools recognize that privacy alone is not enough and that end users must be allowed to make the final decision in these matters.

application allows users to view algorithmic profiles as well as explanations for why they were assigned a particular class. In addition to viewing these explanations, users are able to edit individual tweets locally to see how their algorithmically generated profiles change. They are aided in this process by several tools, including automatic synonym suggestion and a style translation from the current class to the target class. To my knowledge, no other tools which allow users to see their profiles and attempt to change them in real time are available and no research has been conducted into how users opt to alter social media data to obscure algorithmic profiles.

To test this prototype, I conduct a series of lab sessions where participants are provided with the Mockingbird prototype and encouraged to explore their various profiles. I then analyze results from these sessions to distill major themes, draw conclusions about user experiences with algorithmic profiles, and suggest avenues of future work.

Through this research, I hope to address a series of research questions, outlined in the next section.

1.3 Research Questions

This research focuses on a few main questions related to user perceptions of algorithmic profiles, user response to explanations of these profiles, and how users opt to control algorithmic profiles. Specifically, the key questions⁷ of this research are:

RQ1 How important are algorithmic profiles to social media users?

RQ1.1 Do algorithmic profiles meet user expectations?

RQ1.2 Which types of profiles do users react most strongly to?

RQ1.3 Does this importance change after users see examples of profiles with explanations?

RQ2 How do users respond to explanations of algorithmic profiles?

RQ2.1 Do explanations increase or decrease trust in algorithmic profiles?

RQ2.2 Do explanations help users to build mental models of how profilers work?

RQ3 How do users opt to change their social media data in response to algorithmic profiles?

RQ3.1 Do users care about only the final classification or also the confidence?

⁷For clarity, questions are divided into three broad questions, each with more detailed sub-questions, and identified by a bolded tag **RQ1.1**. For sub-questions, hypotheses are offered below with a similar tagging system. **H1.1a** represents the first hypothesis regarding **RQ1.1**, for example.

RQ3.2 What compromises are users willing or unwilling to make to change algorithmic profiles?

RQ1 is meant to address the relevance of studying social media profiles. Existing research has shown that user response to profiles can be quite mixed and complex, weighing both potential upsides and downsides [Ur et al., 2012]. I predict that expectations will be altered to be more moderate, as at least one profile is likely to be more or less accurate than users expect (**H1.1**). I expect that users will care more about personal attributes (such as personality) than text characteristics (such as sentiment) (**H1.2a**) and will focus on profiles that do not match their expectations (**H1.2b**). It is also likely that priorities will differ between users based on both profile accuracy and user preferences (**H1.2c**). Regarding explanations, I think that they will cause users to be less worried about algorithmic profiles (**H1.3**). Although the opposite is likely if users are highly motivated to control perception of at least one attribute, in general I think that seeing flawed profiles will cause people to be less concerned about most attributes [Eslami et al., 2018]. This may not be the case for “hyper-personal” attributes such as Big Five, especially because these attributes do not have an explanation and users may not have a clear “ground truth” or even intuition for how they should be classified before seeing their profile [Warshaw et al., 2015].

RQ2 focuses specifically on the effect of explanations on perceptions of algorithmic profiles. Given that explanations offered by social media sites tend to be insufficient, if not misleading [Andreou et al., 2018], these explanations are likely to be the first time users have insight into how they are profiled online. As outlined above, I believe that in general explanations will cause a decrease in user trust (**H2.1**). Inaccurate profiles with explanations may lead to “algorithm disillusionment” [Eslami et al., 2018], where users find problems with algorithms and this decreases trust. I predict that explanations will influence participants’ mental models but that these models will remain somewhat inaccurate (**H2.2**). [Eslami et al., 2018] found that users attempt to justify algorithmic decisions, even if they are incorrect, by constructing mental models or “folk theories” [Eslami et al., 2016]. These mental models are likely to influence how users edit tweets. For lexical models, this may be a valid method. However, users are not likely to be fully informed and more complex models may not be so easily approximated. In the gender lexicon used for this experiment [Sap et al., 2014], “all” has a weight of -5.95, indicating that it is quite masculine. A misspelling of it (“alllll”) has a weight of 131.43 which is extremely feminine. If a user were to see the misspelling as an explanation for why a tweet was labeled feminine, they may avoid using the correct spelling when in fact it has a masculine weight. Non-linear models are likely to obscure even more complex relationships that cannot be easily determined by looking at a few examples. For this reason,

relatively simple explanations can convince users that they understand complex models even if they do not.⁸

Finally, **RQ3** asks how users change their social media data in response to algorithmic profiles. Each profile has varying degrees of granularity. At the highest level is a simple classification and at the lowest level is the specific score or confidence assigned for that class. Broadly, the tools provided allow for manipulation at any degree of granularity although small adjustments may be more difficult to achieve in practice. I suspect that participants will care about being within certain confidence ranges (**H3.1**). The tools used in this experiment suggest three levels of confidence (“low,” “moderate,” and “high”) and I think that is how most users will frame their target confidences.⁹¹⁰ **RQ3.2** is perhaps the most interesting as well as the hardest to measure. Users who undertake altering their social media data must ensure the edited account remains true to their personality and does not compromise on their most important issues. Additionally, changing data may influence several profiles at once, possibly forcing users to prioritize some profiles over others. In general, I expect users to make relatively few changes the meaning or tone of a social media post but to be more open to making changes which do not alter the meaning of the original post (**H3.2**).

⁸This is prediction is similar to concerns raised in [Gilpin et al., 2018].

⁹This may be something of a self-fulfilling prophecy, in that providing these categories encourage users to think this way. However, users are also free to ignore confidence levels or focus directly on the raw numbers. These confidence levels are suggested to explicitly provide another level of granularity.

¹⁰Note that for lexicons, the raw score is actually a measure of extremity (exactly how old or masculine the text looks) but it is somewhat more complicated for other classifiers. It is convenient to consider class confidence as a proxy for extremity but the two concepts are somewhat distinct - extreme positions are best represented by separate classes, as in the categorical income data classifier used in this study. Even so, users would probably like to control on a certain level the magnitude or confidence of each trait rather than just the final label.

Chapter 2

Background

This chapter provides an overview of several key topics related to this project and places them in context. The first section focuses on explainable and interpretable AI and ML, providing both a general background and a technical explanation of the most relevant explanatory technique. Subsequent sections focus on algorithmically generated profiles based on social media data, existing tools for enhancing user privacy, and a sociological background for why users may desire control over how they are perceived online.

2.1 Explainable & Interpretable AI/ML

This section introduces explainable and interpretable techniques which underlie much of the basis for this project. First, I provide definitions for these terms, noting disagreements in the literature. Next, I consider some open questions regarding explainable and interpretable systems. Finally, I explain in detail the primary explanatory tool used by this paper: LIME [Ribeiro et al., 2016].

2.1.1 Definitions

The literature does not appear to have a consistent definition for either explainable or interpretable models. In fact, provided definitions often conflict, leading to disagreements and confusion. In light of this, it is still possible to make a few general characterizations about their meanings in practice.

Explainable models “summarize the reasons [...] for behavior [...] or produce insights about the causes of their decisions” [Gilpin et al., 2018]. An example of this is a model which returns textual explanations for why it chose a certain action, as in [Ehsan et al., 2019]. Often techniques for explaining models are “post-hoc,” in that they are augmentations to models that are not generally scrutable [Rudin, 2018]. LIME is one such post-hoc tool that explains individual predictions by

approximating the complex model with a simpler interpretable model (see Section 2.1.3). Interpretable models “describe the internals of a system in a way which is understandable to humans” [Gilpin et al., 2018]. An example is a rule list, or a series of “if-else” statements where the conditions are the learned parameters [Rudin, 2018]. The advantage of these models is that it is always easy to trace what the model is doing and identify unintended or inappropriate decision-making factors. Fidelity is also high with interpretable models, in that there can be no doubt that the understanding gained by humans is true to what the model is actually calculating. This is in contrast to explainable models, which often rely on approximations or additional complexity to generate (potentially inaccurate) explanations.

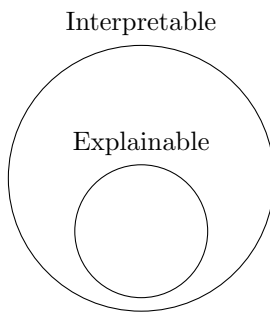


Figure 2.1: Explainability as a subset of interpretability as in [Rudin, 2018].

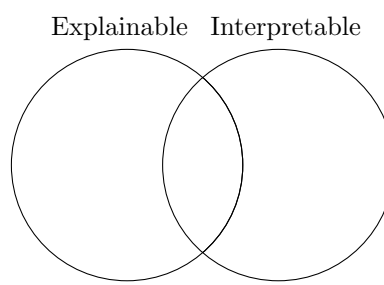


Figure 2.2: Explainability and interpretability as two distinct classes with an intersection.

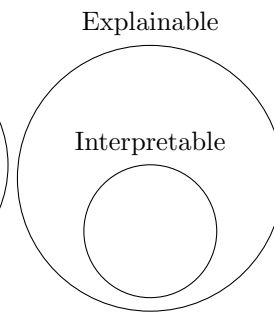


Figure 2.3: Interpretability as a subset of explainability as in [Gilpin et al., 2018].

The relationship between these explainable and interpretable models is unclear - the figures above show three potential relations. “Explainable models are interpretable by default, but the reverse is not always true” according to [Gilpin et al., 2018]. This corresponds to Figure 2.1. In contrast, some authors claim that models which are “inherently interpretable [...] provide their own explanations, which are faithful to what the model actually computes” [Rudin, 2018]. This corresponds to Figure 2.3, where interpretability implies explainability. As there are clearly models which are both interpretable and explainable, this leaves only the possibility of Figure 2.2, which suggests that models can be explainable, interpretable, or both. This option seems most reasonable. Explainable models that use post-hoc explanations such as LIME or learn to generate text-based rationale [Ehsan et al., 2019] are not interpretable in the sense that they are not “constrained in model form” [Rudin, 2018] and do not necessarily offer a true understanding of the way in which decisions are made in the original model. On the other hand, clearly interpretable models such as rule lists [Angelino et al., 2017] may provide rules that seem arbitrary or illogical to humans trying to understand them and so do not explain to a point “when you can no longer keep asking why” [Gilpin et al., 2018]. The intersection of explainable and interpretable models might be something like a rule list which also offers “prototypes” - examples in the training set which contributed most

to each rule [Rudin, 2018]. Although it is not the goal of this paper to draw strict conceptual lines,¹ the above points are meant to roughly illustrate the conceptual framing and assumptions made throughout this paper.

2.1.2 Problems & Open Questions

There are several problems and open questions with regards to explainable and interpretable models and systems. The first question, raised vocally in [Rudin, 2018], is whether or not explainability is even a worthwhile characteristic without interpretability. Rudin argues that “explanations must be wrong” because they “cannot have perfect fidelity with respect to the original model” [Rudin, 2018]. She contends that no explanation system can be trustworthy and that adding complexity to explain the existing complexity of an inscrutable model is problematic. Relying on untrustworthy explanations is likely worse than having no explanations as users are less likely to be skeptical of a model they think they understand. If programmers and testers see explanations that align with their prior assumptions and that do not seem unethical, they may be likely to do less rigorous testing than they would otherwise. Additionally, they may be incentivized towards adopting explanations which put the model in the best possible light or “*persuasive* systems rather than transparent systems” [Gilpin et al., 2018]. As shown in [Binns et al., 2018], the nature of an explanation can influence perceptions of fairness and justice. Seemingly good explanations could make flawed models appear better than they are just as poor explanations could be detrimental to otherwise desirable models.

Rudin advocates for interpretable models instead. However, other authors argue that “interpretability alone is insufficient” and that models need “the capacity to defend their actions, provide relevant responses to questions, and to be audited” [Gilpin et al., 2018]. Some of this disagreement may be attributed to different conceptual framings, as pointed out above. For instance, [Gilpin et al., 2018] consider attention networks in deep learning as “explanation-producing systems” whereas Rudin would likely consider this an interpretable element as it provides direct insight on the calculations of the model. Regardless, the point made by both [Rudin, 2018] and [Gilpin et al., 2018] is that explanations have the potential to be deeply flawed and so careful consideration must be given to their use.

Another open question concerns the appropriate use cases of scrutable models and systems. Interpretable models are not suitable for all cases because they reveal how the algorithm works and leave it open to exploitation and imitation. A rule-based interpretable model could, for

¹In fact, such lines may not be desirable; [Rudin, 2018] argues that “interpretability is a domain-specific notion” and so any general purpose definition may be inherently fraught.

instance, encourage people to lie on resumes to guarantee that they will get a job. The open nature of interpretable models also makes them somewhat unappealing to industry, as proprietary algorithms immediately become public knowledge [Rudin, 2018]. This could be partially mitigated by revealing the interpretable model only to regulators or overseers and treating the public-facing interface as a black box. This would naturally raise its own problems about which groups are permitted to view the inner workings of the model. Government could intervene in certain cases, ensuring that in cases of public interest companies are compensated properly to produce accurate, interpretable algorithms.

Furthermore, interpretable models are often simpler than existing inscrutable models. Deep learning models with millions of parameters are not comprehensible by humans, but replacing them with small interpretable models may come at a serious performance cost. Simpler models may not be able to capture the same information as complex models. [Rudin, 2018] argues that this trade-off is not necessary for most relevant cases and that insights gained from interpretability would overcome diminished accuracy by improving future iterations of the model. This may be domain-dependant, in that simple specialized models may compete with more general and complex ones in high importance applications, or that the decrease in accuracy is deemed less important than the improved clarity.

However, this highlights another problem: interpretable models are likely to require more domain expertise than large deep learning models. One appeal of these general purpose deep learning models is that relevant features and relations can be discovered by the algorithm without extensive data manipulation or “baking in” of expectations based on domain knowledge. Interpretable models will not often have this luxury. Again, in cases where interpretable models are necessary, it will likely be worth the effort and resources to understand the complexities of the domain and tailor a model to it [Rudin, 2018]. It would also be possible to create a feedback loop whereby inscrutable deep learning models are used at first to gain insight into relevant features or problems with the data. Post-hoc explanations could be added for further clarity. These insights could then be applied to building an interpretable model. Shortcomings of this model could be used to revise the deep learning model and this process could repeat iteratively until the interpretable model performs adequately. This complex and costly process may only be justified in a few applications.

With these potential shortcomings, care must be given to select appropriate uses of scrutable algorithms. Even if a use case is chosen, the type of explanation or interpretable model must be determined. For instance, an explanation which relies on prototypes may be useful for programmers to correct anomalies but may leave end users with a sense of unfairness [Binns et al., 2018]. Interpretable models may increase trust to the point that users trust them beyond their own judgment and cease to

be able to recognize when they make mistakes [Poursabzi-Sangdeh et al., 2018, Stumpf et al., 2016]. The way a model is presented is likely to encourage users to be more or less trusting of it and so research in to perceptions is likely necessary before deployment of such models.

Perhaps the broadest open question is of how best to use explainable and interpretable models to increase user understanding and autonomy. The possible applications of these techniques are quite large, however they may face technical, normative, regulatory, and economic challenges. Additionally, barring major legislation, it is unlikely that even effective models would quickly replace entrenched existing practices. The exact application also matters: bad explanations could decrease trust in a “good” system because they misrepresent or poorly explain decisions. Conversely, systems which explain too well can result in too much trust and an unwillingness to question the result of the system [Stumpf et al., 2016]. This question leaves open a wide range of applications for explainable and interpretable models to improve experiences and autonomy of individual users.

2.1.3 Explanations Through Approximation

Most of the focus on model scrutability is centered on design – that is, how do we design models and systems which are scrutable? This has two major limitations: waiting on scrutable models does nothing to help people who desire explanations for existing systems and there is no guarantee that scrutable models will be adopted at all. With this awareness in mind, there has been research into “post-hoc” explanations. These are generally model-agnostic tools which aim to explain decisions made by machine learning algorithms. The most well known of these is called the Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016]. LIME attempts to locally estimate an arbitrary classifier with a linear approximation in order to explain a single classification instance. LIME explanations aim to be both interpretable, in that they are easy to understand by humans, and maintain “local fidelity” [Ribeiro et al., 2016]. An explanation with local fidelity truly explains how the model behaves with respect to a specific instance. Explanations with local fidelity need not be accurate or informative for all possible inputs; the goal is only to provide an explanation for why a specific input was classified the way it was. This is especially important as it allows LIME to treat the classifier as a black box and still generate useful instance-specific explanations.

An explanation $\xi(x)$ for the classification of example x by model f is found according to the following formula:

$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (2.1)$$

where G is the class of interpretable models, π_x is a measure of locality around x , \mathcal{L} is a measure of how accurately g approximates f around π_x , and $\Omega(g)$ is a measure of the complexity of g

[Ribeiro et al., 2016]. As no assumptions are made about f , $\mathcal{L}(f, g, \pi_x)$ is estimated by perturbing x uniformly at random and weighting the importance of each perturbed point by its closeness to x as defined by π_x . This weighting ensures local fidelity. [Ribeiro et al., 2016] restrict g , π_x , and \mathcal{L} to the following forms:

$$g(z) = w_g \cdot z \quad (2.2)$$

$$\pi_x(z) = e^{\frac{-D(x,z)^2}{\sigma^2}} \quad (2.3)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) (f(z) - g(z'))^2 \quad (2.4)$$

Here D is a distance function with width σ and w_g is a vector of learned weights. Model complexity $\Omega(g)$ depends on the class of g .²

The LIME approach is somewhat limited, as the class of interpretable models G considered is substantially restricted. Additionally, the explanations come from approximations rather than a true understanding of the underlying model. For this reason, when used for debugging LIME cannot do more than point in the direction of potential problems in the original model. LIME also tends to be slow, even for short text examples but especially for images, as it relies on a large number of random perturbations. This makes it ill-suited to instantaneous customer feedback and limited to evaluating a few selected examples.³

[Ribeiro et al., 2016] found that LIME helps users remove irrelevant features and identify problematic models even if those users had little to no background in machine learning. This is partially a result of having accurate explanations and is not specific to the LIME,⁴ although [Ribeiro et al., 2016] also found that LIME substantially outperformed various other explanatory methods in identifying important features. Despite its shortcomings, LIME is certainly a valuable tool to generate explanations for pre-existing black box models.

2.2 Building Profiles from Social Media

This section introduces profiles based on social media data. First, I motivate the usage of social media data. Next, I outline several of the attributes which are most frequently profiled based on social media and several profiling techniques which are used. Then I describe in detail a profiling

²In the case of text classification it is given by 0 if fewer than K words are used in the explanation and infinity otherwise [Ribeiro et al., 2016].

³Overcoming this limitation to understanding the entire underlying model is addressed in [Ribeiro et al., 2016], where a system for selecting representative instances to understand the model more broadly is presented.

⁴Any good explanation should help with these problems.

network used in this project which relies on features of social media accounts other than the text of posts. Finally, I consider how users respond to algorithmically generated profiles based on their social media data.

2.2.1 Use of Social Media Data

In recent years, much ink has been spilled over the power and influence of the Internet on individual decisions. Social media has been scrutinized in particular, both as a platform designed to manipulate users and as a vehicle for echo chambers. Recommendation algorithms that sort newsfeeds and suggest new connections are a feature of daily life for most Internet users but it is not difficult to see how such algorithms could be manipulated to encourage specific behaviour. The risk of such manipulation is increased when users do not understand how the algorithms work or place too much trust in them. Publicly available data on social media networks can often be quite telling of user characteristics and may be used to “profile” users without their knowledge or consent. Users may not even be aware such profiling is possible from the data they have made available, let alone how it may influence the content they see. These profiles are in turn used in a variety of discrimination tasks. Companies like Facebook and Twitter use these profiles (often along with third-party data) to sell targeted ads based on characteristics such as age, gender, income, and political affiliation [Andreou et al., 2018]. Although public backlash has forced Facebook to reform their ad targeting practices, advertisers used to be able to filter on race [Angwin and Parris Jr, 2019] or interest in hate groups [Angwin et al., 2019].⁵ As the selling of ads is the primary source of revenue for these corporations, and targeted advertising tends to work better than untargeted [Matz et al., 2017], there is a substantial incentive for corporations to build and use these profiles. There is also an incentive for corporations to keep these profiles and the methods by which they are generated secret, as the algorithmic creation of profiles becomes the “secret sauce” - a competitive advantage in selling ads.

Clearly, such automatically-generated profiles can be detrimental to the average user of social media, especially if the profiles are on sensitive attributes or are inaccurate. Consider the following cases where an algorithmically generated profile can have a negative effect on an individual by revealing sensitive information:

- A closeted member of the LGBTQ+ community constantly receives targeted advertisements for products or activities associated with that community. These ads may be seen by coworkers,

⁵In some cases, the root cause of racially biased advertising may be harder to identify. In [Sweeney, 2013], it is unclear whether ad providers, Google, or “society” encouraged Google to serve significantly more ads involving the word “arrest” to searches on traditionally black names than on traditionally white names.

friends, and family members incidentally, depriving the individual of the power to decide who they come out to.

- A woman attempting to hide a pregnancy is targeted with maternity ads. In one case, such an automated ad profile by the American supermarket Target resulted in a father discovering his teenage daughter’s pregnancy.⁶
- A user is mistakenly classified as having interest in a hate group or other undesirable community and is bombarded with posts related to it.

These effects may be even worse if the data are used to profile for the purposes of job recruitment [Smith and Kidder, 2010], loan applications, or dating profiles. [Ali et al., 2019] showed that Facebook may skew ad delivery along gender or racial lines even if advertisers do not select for these characteristics explicitly. This is clearly problematic for advertisements around job openings or housing, as these topics are legally required to refrain from discrimination along racial or gender lines.

Social media profiling is also problematic because even if users are aware that they are being profiled, they are often unable to see what their algorithmically generated profile is, let alone change it. They are also unable to determine how their algorithmically generated profiles affect their lives or user experience. [Andreou et al., 2018] showed that explanations of ads offered by Facebook are often misleading and vague. Furthermore, they often offer little in the way of actionable change. It is easy to hide advertisements or report them, but these address the symptom (seeing the ad) rather than the cause (Facebook’s profile of you). [Andreou et al., 2018] also found that users can be targeted based on “hidden attributes,” that is characteristics that are not available for users to control in their ad preferences. These hidden attributes may be based on data from third party data brokers, which users have even less control over and knowledge of. Facebook and Twitter both offer options to remove automatically generated interest tags, but it is unclear how often these results are updated. These controls also omit “hidden attributes,” which may be more relevant to users than interest in a particular topic. Lastly, little to no insight on why a tag was generated in the first place or how to prevent such tags from being generated in the future is provided.

Most research into algorithmically generated profiles has focused on the social media sites of Twitter [Wang et al., 2019, Pennacchiotti and Popescu, 2011, Gou et al., 2014], Facebook [Kosinski et al., 2013b], and Reddit [Chen et al., 2014]. A meta-analysis found that Facebook and Twitter dominated the research on Big Five traits [Azucar et al., 2018]; this is likely due to the two

⁶<https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html> [Duhigg, 2012]

sites' prominence in social media and ease of attaining data.⁷ A handful of profiling methods focus exclusively on user profile data [Wang et al., 2019] or network analysis [Golbeck and Hansen, 2011], but most include an analysis of text (tweets or Facebook posts).

2.2.2 Types of Profiles

Algorithmically generated profiles for social media users may be quite complex, but tend to focus on a few characteristics. The simplest profiles build interest tags from things such as page likes: characteristics users have explicitly provided to the network. These are often the most intuitive and tend to offer users the most control. These profiles are often secondary to more general information, which can be inferred algorithmically or provided by a third-party data broker. General demographic profiles include characteristics such as age, gender, education level, relationship status, political affiliation, and income [Dewey, 2016].

Basic Demographic Data

Basic demographic information, such as age, gender,⁸ income, and location, appears to be some of the most important for ad targeting. To give an example, Experian, a credit reporting company, offers a marketing program called Mosaic which allows companies to appeal to users based on age, location, income, and a few other characteristics.⁹ Facebook also uses these characteristics, frequently including age, gender, and location in explanations for ads [Andreou et al., 2018]. Although age, gender, and location are likely to be provided by users when they join Facebook, they may be less apparent on other forms of social media.

A common approach to profiling age and gender is to build lexica [Sap et al., 2014] tailored to social media (specifically Twitter). Lexical classification methods focus on social media posts, learning words which correspond to each class. Lexicons are among the simplest and most intuitive way of classifying texts, and, once built, are easy and efficient to use.

In general, a lexicon l classifies a document d according to the following equation:

$$y_l = k + \sum_{t \in d} \frac{l(t) * \text{count}(t, d)}{|d|} \quad (2.5)$$

⁷Both sites have recently made it more difficult to share data. Many studies mentioned in [Azucar et al., 2018] used the MyPersonality dataset from Facebook which is no longer available to researchers. Twitter still provides an API but has since restricted its usage and altered its terms of service to prevent sharing the of existing datasets.

⁸The term “gender” here and in the existing literature is treated as a binary characteristic and so is not inclusive of all gender identities. This shortcoming is likely to be present in industry. As this project is meant to build upon existing industry practices, it does not seek to directly address this issue, even though it is an important one deserving of further study. Instead, this project aims to empower users to control how their gender identity is perceived online. Although there is no explicit non-binary option, a prediction of gender with little or no confidence may be a reasonable proxy - importantly, it is up to users to make this decision.

⁹<https://www.experian.co.uk/assets/marketing-services/brochures/mosaic.uk.brochure.pdf>

where each t represents a token in d , k is a constant intercept term, $\text{count}(t, d)$ is the number of times token t appears in document d , $|d|$ is the number of tokens in document d , and $l(t)$ is the weight that lexicon l gives to token t or 0 if t is not in l . Here, y_l is shown as a real-valued label but the equation can be generalized to assign d a categorical label based on the value of y_l . For binary classification the sign of y_l is often used to denote the class.

[Sap et al., 2014] observed that this equation looks very similar the one used by linear models for classification and regression. Using a dataset of Facebook posts from over 75,000 users who provided their age and gender, they trained classifiers based on 1-grams from these texts. For age, [Sap et al., 2014] used ridge regression [Hoerl and Kennard, 1970]. They treat gender as a binary classification problem, using an Support Vector Machine (SVM) with a linear kernel and L1 penalization. The weights from these classifiers, along with the intercept, are then used directly as the lexical weights. They achieved a Pearson correlation coefficient of 0.83 for age and an accuracy of 91.9% for gender, and the provided lexicons have been used in other research (e.g. [Preoŧiuc-Pietro et al., 2015b]).

Lexical classification is also quite scrutable: it is trivial to see what words caused the classification. However, lexicons are brittle and not able to detect misspellings or learn new relevant words. Recently, some more complex methods have been used to predict age and gender. [Wang et al., 2019] used a multimodal model incorporating several neural networks to classify users based solely on their profile data. Although this model is significantly more complex, it achieves state-of-the-art accuracy and is able to incorporate profile images. This is achieved by sacrificing scrutability, as the model is not interpretable and offers no explanations.

While location can often be gathered directly from user profiles or IP address, some research has focused on deriving this information purely from tweet content. [Zheng et al., 2018] outline several common methods, including identifying words specific to a certain locale, word-based machine learning techniques, and network-based inference. These approaches face a number of challenges, including sparsity for particular regions and identifying meaningful locations for users who frequently move between locations.¹⁰ Frequently locations are constrained to be major cities, for which there is a significant amount of data and beyond which more granular profiling may not be necessary. Location information can be particularly important in the case of disasters or predicting local electoral results.

There have also been attempts to predict user income based on social media data. In one project, [Preoŧiuc-Pietro et al., 2015a] analyzed about 5200 Twitter accounts classified into UK

¹⁰For instance an international or out-of-state college student. In this case, the most desirable location likely depends on the application.

job categories based on information provided in user biographies (bios). Using profile metadata along with word embeddings, they trained a Gaussian Process (GP) [Rasmussen, 2003] as well as an SVM to predict nine job categories and achieved 52.7% classification accuracy. They also found that user-level features such as number of followers were not useful in classification. The authors attempted regression on the same dataset [Preoțiuc-Pietro et al., 2015b], using the mean income for each occupation as an estimate for each user. The Mean Average Error for the best classifier was £9,535, indicating a relatively accurate model.¹¹

Big Five and Other Psychological Profiles

Psychologists have developed several ways of classifying characteristics of an individual’s personality and studied these classes to describe or predict individual behavior, well-being, and success [Ozer and Benet-Martinez, 2006]. Although such classes inherently simplify and diminish individual factors or circumstances, they are an active area of study and often used to describe basic traits of people. One of the most popular personality classification systems is known as the Big Five [McCrae and John, 1992]. This system generally uses a questionnaire to score users across five categories: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Higher scores indicate a tendency towards the highlighted characteristic, whereas lower scores indicate a tendency towards its opposite.¹² Big Five characteristics have been shown to correspond with social media use and so have proven a popular area of research for constructing profiles from social media data [Azucar et al., 2018].

Profiles using Big Five characteristics have come under intense scrutiny since reports on the company Cambridge Analytica came to light. Cambridge Analytica gathered information from Facebook profiles, often without consent, and used that information to build psychological profiles. These profiles were then used to create highly-targeted ads in US political campaigns [Cadwalladr and Graham-Harrison, 2018]. Profiles relied on data such as Facebook page likes as in [Kosinski et al., 2013a] and [Youyou et al., 2015]. While it is ultimately unclear how effective or widely used these models were [Confessore and Hakim, 2017], their existence caused a public backlash, triggered several investigations, and prompted social media platforms to restrict access to their data. The incident has also sparked a public curiosity about how such automated profiling tools work.

While Big Five is one of the most prevalent psychological profiles,¹³ research has also focused on

¹¹Note that this analysis is based purely on average income and is unlikely to be robust to atypical jobs and situations. It also fails to capture any sort of existing wealth, property, or additional income provided by investments or by a spouse or partner.

¹²For instance, a high score for Extraversion would suggest someone is outgoing and prefers large parties, whereas a low score would indicate they are more introverted and prefer quieter, smaller settings.

¹³Studied in, for instance, [Mairesse et al., 2007, Golbeck et al., 2011, Quercia et al., 2011, Kosinski et al., 2013a,

other personality profiles. In addition to Big Five, [Gou et al., 2014] considers “fundamental needs” and Basic Human Values, each studied individually by [Yang and Li, 2013] and [Chen et al., 2014] respectively. It is worth noting that many values and needs overlap with Big Five categories and there is little reason to believe that they are more or less effective than Big Five in modelling personality.

For textual analysis, most profiling tools use Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2015]. The LIWC application scans text, looking for key words in its lexicon. Each word in the lexicon belongs to one or more categories. Categories may correspond to parts of speech, verb tenses, punctuation, emotional states, or topics like death and money. LIWC returns the percentage of words in the text that belong to each category. Various authors have found significant correlations between LIWC category frequency and personality profile score [Mairesse et al., 2007, Golbeck et al., 2011, Gou et al., 2014, Chen et al., 2014]. Some researchers [Farnadi et al., 2016, Mairesse et al., 2007, Golbeck et al., 2011] also use the Medical Research Council Psycholinguistic Database (MRC) [Coltheart, 1981] which works in a way similar to LIWC but with different categories. Regression is performed over these linguistic features to generate the integer personality profile scores [Chen et al., 2014, Farnadi et al., 2016]. Some research has shown that decision trees can outperform SVMs for this classification [Farnadi et al., 2016]; decision trees also have the desirable property of interpretability. Predictions averaging within 11% of the true score generated by taking a survey have been achieved [Golbeck et al., 2011] using these methods.

Some recent work has moved away from closed vocabulary approaches such as LIWC.¹⁴ [Arnoux et al., 2017] present a method to classify users on Big Five characteristics with little input data by combining GloVe embeddings [Pennington et al., 2014] with a GP. Each tweet is represented as an average of the GloVe embeddings of the words it contains. These representations are used to train a GP. Training on up to 200 tweets from 1000 users, [Arnoux et al., 2017] found that the GP outperformed other models, including LIWC-based regression, by up to 37% and required less data than these other methods. It is worth noting that these regression methods are still limited in terms of accuracy - [Arnoux et al., 2017] report an average correlation of 0.33, with a minimum of 0.25 for extraversion. The GP was then incorporated into IBM’s public facing personality tool which provides Big Five, needs, and values profiles based on Twitter or text data.¹⁵

Gou et al., 2014, Youyou et al., 2015, Farnadi et al., 2016, Azucar et al., 2018]

¹⁴It is worth noting that the LIWC tool is not open source and that it costs money to use the service.

¹⁵<https://personality-insights-demo.ng.bluemix.net/>

Political Affiliation

Another common application of social media profiling is political affiliation. Understanding the political affiliation of a user is crucial for the study of echo chambers and online communities in general. Most research treats this as a binary classification problem, labelling users as broadly conservative (right) or liberal (left) [Conover et al., 2011, Golbeck and Hansen, 2011]. As much research is centered on US politics, the categories of Republican and Democrat are also common [Pennacchiotti and Popescu, 2011, Kosinski et al., 2013a]. Classification can be difficult as “conservative” and “liberal” are somewhat nebulous and situationally-dependant labels. Additionally, associating a user with a political ideology does not mean that they agree with all planks of the party platform. Research has discovered a tendency to use parody or sarcasm to mock the other side as noted in [Conover et al., 2011]; this may result in people using hashtags, phrases, or language that is common among people who have beliefs that are opposed to their own. Network analysis is most prevalent for identification of political affiliation, as it is supposed that users tend to follow politicians whose views they agree with [Golbeck and Hansen, 2011] or connect with users who have similar views. In [Conover et al., 2011], researchers were able to identify political affiliation with 95% accuracy by simply clustering users into two groups based on network topology.

Other Characteristics

Although the above characteristics are some of the most frequently studied, they are not the only profiled characteristics. [Wang et al., 2019] also profiles based on “organization status,” which is meant to indicate if a Twitter account is corporate or private. The organization-status label may be particularly useful for filtering accounts not run by individuals sharing their own views. Others have worked on identifying “bots” on Twitter [Dickerson et al., 2014], that is, separating accounts controlled by humans directly from those running automated programs.

A common problem in Natural Language Processing (NLP) is sentiment analysis: determining whether a given piece of text expresses an opinion that is broadly negative or broadly positive. Sentiment analysis may also be an interesting feature in social media profiles; [Dickerson et al., 2014] use it as main feature for identifying bots. They found sentiment features to improve accuracy by about 53% relative to classifiers lacking sentiment features, and that human users tend to have more extreme sentiment than bots. It is not difficult to imagine using sentiment as a proxy for how users are likely to respond to “positive” or “negative” advertisements. Sentiment analysis can be approached through lexicons trained on Twitter data [Kiritchenko et al., 2014, Mohammad et al., 2013, Zhu et al., 2014], simple classifiers such as SVMs and Naive Bayes Classi-

fiers (NBC) [Pang et al., 2002], and complex deep learning methods [Dos Santos and Gatti, 2014].

Other authors have looked at subjectivity, ethnicity, education [Culotta et al., 2015], sexual orientation, religion, and substance use [Kosinski et al., 2013a]. It is not hard to see that these characteristics may be quite private for some users and that creating profiles on them without consent could constitute a serious breach of trust. Even if users consent to be profiled on these data, it may be necessary to ensure that the profiles do not cause harassment or other harm.

2.2.3 Classification From Profile Data (M3)

While most of the other classifiers mentioned above take any text as input, others focus on different aspects of social media profiles. One such model is developed in [Wang et al., 2019].

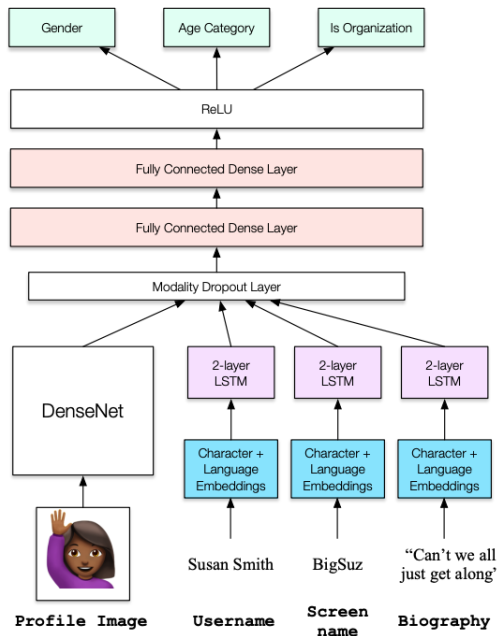


Figure 2.4: M3 architecture as presented in [Wang et al., 2019].

[Wang et al., 2019] aim to identify the age, gender, and organizational status of users to facilitate a better demographic understanding of Twitter data and propose a multimodal, multilingual, and multi-attribute (M3) model to do so. The M3 network works by learning a representation for each of the four input fields (user name, user tag, profile picture, and bio), passing these as input to a dense feed-forward network, and then using separate dense layers to get the probability distribution for each class (see Figure 2.4). To construct this network, separate text and image classifiers are trained. The image classifier makes use of DenseNet [Huang et al., 2017], a state-of-the-art image classification model. DenseNet improves existing convolutional feed-forward networks by adding direct connections from all layers to all subsequent layers via concatenation. This reduces

the number of model parameters as well as improving information flow [Huang et al., 2017]. The text classifiers begin with character embeddings concatenated with trainable language embeddings. These embeddings are passed to a dense layer and then a pair of bi-directional LSTM layers. Using character-level embeddings reduces the vocabulary size when compared to a word-level embedding and aids classification in the multi-lingual setting of Twitter [Wang et al., 2019].

After being trained separately, the output layer of these models is removed and they are combined into the M3 network. The M3 network consists of these inputs, a modality dropout layer, two dense layers with 2048 units each, a ReLU activation [Nair and Hinton, 2010], and then separate dense layers for each attribute with softmax activation functions. The complete network is initialized with the weights of the separately trained models and is then trained again end-to-end. The modality dropout layer randomly removes some of the four input layers to improve regularization and prepare the model for missing data; each input is dropped with probability 0.25. Dropped images are replaced with a random matrix and dropped text is replaced with an empty embedding. By using co-training and automatic translation to English on training data, [Wang et al., 2019] further augment the training process.

[Wang et al., 2019] found the M3 network outperforms all tested models on Twitter data and performs at least comparably on non-Twitter data for all three attributes.¹⁶

2.2.4 User Response

As argued above, few users are aware of their algorithmically generated profiles and even fewer have had the opportunity to view them. When informed about their profiles, opinions are often mixed. In the case of online behavioral advertising, users acknowledge that online profiles are likely to improve the relevancy of their ads but voice privacy concerns [Ur et al., 2012]. Users are comfortable being tracked on some characteristics (such as news consumption) and uncomfortable being tracked on others (searches pertaining to sexually transmitted diseases) and more comfortable being tracked by Google than other companies [Ur et al., 2012].¹⁷ Many users observed errors in their profiles, although few were concerned about this or tried to correct the errors [Rao et al., 2015]. This may reflect the purpose of the profiling: it is not always obvious that ad profiles can be damaging to users. It is reasonable to expect a desire for more control in higher-stakes applications.

In cases where profiles are more subjective, users were hesitant to change “expert” algorithmically generated profiles [Warshaw et al., 2015, Eslami et al., 2018]. This trend sometimes reversed

¹⁶The M3 network was trained on about 14.5 million examples for gender, 2.6 million for age, and 23.9 million for organizational status.

¹⁷This paper was published in 2012, so it is possible that user attitudes towards particular companies have shifted since the interviews were conducted.

when profiles were shown to be inaccurate, a phenomenon referred to as “algorithm disillusionment” as observed in [Eslami et al., 2018]. As participants saw more and more errors, over 80% realized that algorithms were not as accurate as they had initially thought and many “started to confront their algorithmic self” [Eslami et al., 2018]. Participants also expressed concern about being profiled without their consent or their profiles being provided to certain decision makers. However, participants felt pressure to share their profiles both for the positive benefits of improved recommendations and because they feared that not sharing would be perceived as hiding something negative [Warshaw et al., 2015]. Interestingly, some users blamed their own decisions for algorithmic failures, stating that they had purposefully withheld information that may have been useful for profile generation [Eslami et al., 2018].

2.3 Privacy Enhancing Technology

This section provides an introduction to several challenges and some existing work in privacy enhancing technology. It also suggests why privacy alone may be insufficient and explains in detail one privacy-enhancing tool used directly in this project: A⁴NT [Shetty et al., 2018].

Existing literature regarding changing user identity online tends to focus on privacy rather than empowering users to control their profiles in a more nuanced manner. This distinction is important for several reasons. Firstly, automated privacy tools push edited texts to follow a certain distribution, for instance one where pre-trained classifiers are unable to outperform random guessing. This is the goal of [Shetty et al., 2018], wherein the authors use a Generative Adversarial Network (GAN) [Goodfellow et al., 2014a] to build models that “translate” text from one class to another (see Section 2.3). Even if this tool is successful in fooling complex classifiers and generates semantically correct text, it does not allow users to make decisions about how they are willing to compromise their original text for the sake of privacy. Such models are still quite powerful as they are able to generate replacement sentences for users and when these are treated as suggestions rather than as an automatic substitution, they provide the benefit of enhancing privacy with the nuance of an approach more focused on user control.

Secondly, it may be reasonable to expect that a user does not only care about privacy or anonymity but also the degree to which they are perceived to be a certain way. If users consider algorithmically generated profiles as a proxy for how humans perceive them on social media, they may want to use these tools to tweak perceptions rather than completely confuse them. For instance, if a user is profiled as “highly liberal” or “liberal” with high confidence but sees themselves as more moderate, they may desire to simply decrease the certainty of the classifier without changing the

class they are assigned. Individuals do not fall neatly into simple categories and offering them this level of control gives them more autonomy regarding how they are perceived. This control does not preclude the possibility of users completely embracing explicitly privacy-enhancing tools so no privacy is unwillingly sacrificed by increasing user autonomy.

Privacy may be becoming a more complex question than ever, as recent research has shown that basic demographic attributes (including many frequently made available on social media or profiled as outlined in Section 2.2.1) can be used to de-anonymize data to a surprising degree. [Rocher et al., 2019] recently showed that fifteen demographic characteristics were sufficient to uniquely identify 99.98% of Massachusetts residents. If a user wants to keep their social media account anonymous, simply changing their profiled gender may be sufficient to disrupt this re-identification.¹⁸

Relevant privacy literature tends to focus on either minimally transforming user text to change the user’s predicted class or obfuscating user characteristics. Often these ideas overlap, but obfuscation generally intends to disguise peculiarities of individual authors, such as particular diction or turns of phrase. The intent is to prevent de-anonymization of authors through comparison to previous known writings. This may be accomplished through either editing profiles to undermine classifier confidence [Emmery et al., 2018, Shetty et al., 2018] or to getting classifiers to assign a target class to the instance [Reddy and Knight, 2016].

The latter of these approaches is an application of adversarial examples [Szegedy et al., 2013, Goodfellow et al., 2014b]. Creating textual adversarial examples is generally difficult because almost all perturbations or alterations of text are visible to its human readers.¹⁹ Language is discrete and has complex semantic restraints which are difficult to meet by adding noise, making it a challenging field for adversarial attacks. Some research has attempted to add errors that keep the text readable to humans but break automated classifiers. [Hosseini et al., 2017] add extra letters, periods, spaces, or misspellings to fool a toxicity filter and [Ebrahimi et al., 2017] develop a white-box adversary which changes, adds, or removes single characters at a time. While these attacks are likely to work in some applications, in others the ungrammatical text may be inappropriate. More complex systems, such as models which seek to fool question answering systems, rely on crowdsourcing to validate the final changes [Jia and Liang, 2017].

To address the problem of ungrammatical or illogical examples, there have been attempts at generating “natural” adversarial examples, that is to say examples that are semantically correct.

¹⁸This assumes that the data being used to de-anonymize is coming from algorithmically generated profiles. This may be an increasingly more reasonable assumption as data protection laws become stronger and more aware of de-anonymization attacks.

¹⁹This is in contrast to adversarial examples in image data, which may be imperceptible to humans, see for instance [Szegedy et al., 2013]. Of course, adversarial examples may use human-perceptible perturbations but such examples are easier to identify and correct.

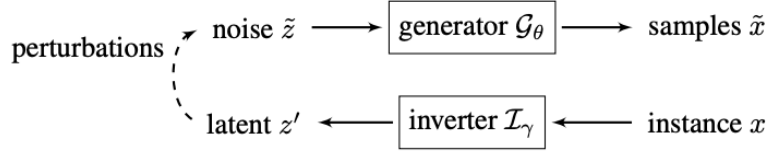


Figure 2.5: Generating natural adversarial examples using the generator-inverter framework. \tilde{x} is chosen as the adversarial example if it minimizes the L_2 norm between z' and \tilde{z} for all \tilde{z} such that $f(x) \neq f(\tilde{x})$ for the attribute classifier f . Diagram from [Zhao et al., 2017].

[Zhao et al., 2017] propose a model which uses a GAN to build a generator-inverter pair. The generator maps random dense vectors to the training distribution while the inverter maps data instances to dense representations. These are trained separately to minimize “reconstruction error,” so that when a dense vector x is passed to the generator and then the inverter, the output is close to x and similarly when a sentence y is passed to the inverter and then the generator, the output is close to the original sentence y [Zhao et al., 2017] (see Figure 2.5). To generate an adversarial example, first the original sentence is passed to the inverter to generate a dense (i.e. lower dimensional) representation. Next, this dense representation is perturbed and mapped by the generator back to the training distribution. The output of the generator is the adversarial example. In training, this is output is used to query the critic and determine if the generator was successful in “fooling” the critic. This approach is “natural” because it locates the adversarial example which is closest to the original sentence in the dense latent space. This latent space is meant to better capture the semantic space of the training data, so that perturbations in the latent space will be more semantically similar to the original instance than perturbations in the input space [Zhao et al., 2017]. For text, the input space could be a range of characters. Randomly changing characters in a sentence is unlikely to result in a new sentence which is both grammatical and semantically similar to the input sentence. However, it is quite likely that, for instance, substituting a word in the original sentence with a synonym will have a small effect on the sentence’s dense representation. The intuition of this approach is that by searching the latent space instead of the input space adversarial examples will be semantically similar to the input sentence. By using human evaluation [Zhao et al., 2017] found that their textual adversarial examples were generally grammatical and semantically similar to the original.

A⁴NT

In a similar project, [Shetty et al., 2018] use GANs to generate adversarial examples. They call their network A⁴NT for Author Adversarial Attribute Anonymizing Neural Translation and aim to use it to re-style text from one class to another. They cast this as a problem similar to that

of neural machine translation (NMT) and so base the A⁴NT network on sequence-to-sequence models used for that task [Sutskever et al., 2014]. To overcome the lack of paired data for training, they use a GAN framework. GANs work by using a generator (in this case the A⁴NT network) to produce artificial samples from the distribution of the training data. A discriminator (in this case the attribute classifier) attempts to separate examples of training data from examples created by the generator. By training the discriminator and generator together, the generator learns to mimic the underlying distribution of the training data [Shetty et al., 2018].

The discriminator is a Long-Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] encoder on word-level embeddings. Each token in the input sentence x is one-hot encoded and fed into a learned embedding layer,²⁰ which is then used as input to the LSTM. The final sentence embedding $E(x)$ is given as the final LSTM output h_{n-1} concatenated with the average of the LSTM outputs from time steps 0 to $n - 2$ where n is the length of x :²¹

$$E(x) = \left[h_{n-1}; \frac{1}{n-1} \sum_{i=0}^{n-2} h_i \right] \quad (2.6)$$

A final linear layer with a softmax activation is used to calculate a probability distribution across classes, and the model is trained using a cross-entropy loss function [Shetty et al., 2018].

The generator, the A⁴NT network, uses an encoder-decoder structure. The encoder is of the same architecture as the discriminator, and outputs $E(x)$ as in Equation (2.6). The decoder pairs $E(x)$ with the previous output word \tilde{w}_{t-1} to generate the next word according to the following equations:

$$h_t^{\text{dec}}(x) = \text{LSTM}[E(x), W_{\text{emb}}(\tilde{w}_{t-1})] \quad (2.7)$$

$$p(\tilde{w}_t|x) = \text{softmax}(W_{\text{dec}}h_t^{\text{dec}}(x)) \quad (2.8)$$

where W_{dec} is a matrix mapping h_t^{dec} to the size of the vocabulary and W_{emb} produces the word embedding. The output word \tilde{w}_t is found by sampling from $p(\tilde{w}_t|x)$. As there is no “true” paired sentence against which we can evaluate the translation, the discriminator is used instead. Texts that are classified as the target class improve the value of the loss function. [Shetty et al., 2018] achieved their best results by training a single encoder for each attribute pair and a separate decoder for each way of transferring style. The task of translating from a male sentence to a female sentence

²⁰[Shetty et al., 2018] state that learning the embeddings from randomly initialized vectors works better than using pre-trained embeddings such as GloVe [Pennington et al., 2014]. They hypothesize this is because the learned embeddings are specific to the attribute are generated.

²¹There is a slight discrepancy here in the equation provided; however, this interpretation seems the most reasonable.

and the task of translating from a female sentence to a male one would use the same encoder but different decoders. This saves training time relative to learning two new encoders but allows the model to learn distinct ways of handling each representation [Shetty et al., 2018].

When translating after training is complete, output tokens are sampled until an ‘END’ token is returned. However, as this sampling is not differentiable it is not suitable for training and so is approximated using the Gumbel-Softmax distribution [Jang et al., 2016]. The Gumbel-Softmax distribution is a continuous, differentiable distribution over categorical data. It has a temperature term τ . When $\tau \rightarrow 0$, the Gumbel-Softmax distribution becomes exactly the categorical distribution; when $\tau \rightarrow \infty$, the Gumbel-Softmax distribution becomes the uniform distribution. The Gumbel-Softmax distribution performs well as an estimator for the categorical distribution, especially when annealing is performed on τ [Jang et al., 2016].

While the attribute classifier is trained with a standard cross entropy loss, the loss function for the A⁴NT network is more complicated. If we define Z_{xy} as an A⁴NT network trained to transfer style from class x to class y , its total loss function L_t is given by:

$$L_t(Z_{xy}) = w_{\text{style}}L_{\text{style}} + w_{\text{sem}}L_{\text{sem}} + w_{\text{lang}}L_{\text{lang}} \quad (2.9)$$

The weights in Equation (2.9) were chosen to make the three terms approximately equal when training began, but the exact values did not appear to be important [Shetty et al., 2018].

If we define s_x as the original sentence, $\tilde{s}_y = Z_{xy}(s_x)$, and $p(\tilde{s}_y)$ as the probability of \tilde{s}_y belonging to class y according to the attribute classifier, we have:

$$L_{\text{style}} = -\log(p(\tilde{s}_y)) \quad (2.10)$$

This is the traditional generator loss function which seeks to maximize the confidence of the attribute classifier on generated examples. To preserve the semantic validity and smoothness of the output text, [Shetty et al., 2018] add L_{sem} and L_{lang} . They propose two potential functions for L_{sem} but find that maximizing reconstruction probability performs best. The loss equation is given by:

$$L_{\text{sem}}(\tilde{s}_y, s_x) = -\log \left(\prod_{t=0}^{n-1} p_{Z_{yx}}(w_t | \tilde{s}_y) \right) \quad (2.11)$$

where $p_{Z_{yx}}(w | \tilde{s}_y)$ is the probability of the network Z_{yx} outputting word w given input \tilde{s}_y and $w_0 \dots w_{n-1}$ are the words which form the original sentence s_x . The intuition behind this function is that if Z_{xy} produces a sentence that can be easily transformed back into the original sentence by Z_{yx} , then little or no information is being lost in these two transformations. This in turn

implies that s_x is semantically similar to s_y . Equation (2.11) represents this “cycle constraint” by calculating the likelihood of Z_{yx} reconstructing s_x given \tilde{s}_y and penalizing sentences from which it is difficult to reconstruct the original [Shetty et al., 2018].

Finally, L_{lang} is meant to enforce “correct grammar and word order” of generated sentences by comparing them to a language model M_y [Shetty et al., 2018]. The authors hope that by choosing generated sentences that are likely to be produced by a language model, the generator will learn to mimic the generally grammatically correct output of language models. The loss is given by:

$$L_{\text{lang}}(\tilde{s}_y) = -\log(M_y(\tilde{s}_y)) \quad (2.12)$$

where $M_y(\tilde{s}_y)$ is the probability that the language model M_y would generate the sentence \tilde{s}_y .

Overall, [Shetty et al., 2018] found this model effective at obfuscation - translations reduced F1 scores to below random chance, both on the adversarially trained LSTM attribute classifier and other baseline classifiers. They were not able to fool classifiers in every instance though, indicating that the A⁴NT network is somewhat limited. Although this performance was promising, translations were not perfect. When compared to sending the input sentence through several rounds of Google Translate, the A⁴NT network output was only preferred about 60% of the time. Additionally, even if the A⁴NT network outperforms Google Translate in terms of semantic similarity, it does not mean that either model is close to the original sentence. Even in cases where the sentence is grammatical and overall similar to the original, the style translation may change certain key words. For instance, [Shetty et al., 2018] includes examples such as “wife” being turned into “crush” or “skool” being changed to “wedding.” Clearly these changes destroy the meaning of the original sentence and could lead to serious confusion when used on social media. This highlights the need for user control to select which words substitutions are appropriate. These translations work on only one attribute at a time, and there is no reason to believe attributes are necessarily orthogonal. If, for instance, there are words or phrases popular with both older users and males, attempting to translate young feminine sentences could result in changes across both attributes. This potential trade-off is one that must be left to user decisions.

2.4 Self Curation - A Sociological Overview

People have managed other’s impressions of them since long before social media was invented. In [Goffman et al., 1978], sociologist Erving Goffman proposes a model of social interaction in which each individual is akin to a performer presenting themselves on different stages. The individual

aims to manage the impressions that members of the audience have of them and is likely to convey different impressions to different audiences - i.e. someone may want to appear hard-working and dedicated in the workplace and easy-going and jovial with their friends. Goffman suggests that individuals try to create impressions such that a “just individual” will treat them “in a way they want to be treated” but that “the observer’s need to rely on representations of things itself creates the possibility of misrepresentation” [Goffman et al., 1978]. This kind of curation or impression management is core to social interaction because it is how people signal to others how they want to be treated. The fact that it is based on indirect representations rather than stated preferences makes it susceptible to confusion and misunderstanding. Goffman also insists that the impressions individuals want to create need not be accurate to their true selves. Indeed, “a sign for the presence of a thing, not being that thing, can be employed, in the absence of it”: people create false impressions or manage Potemkin villages of identity to control their own narrative of self [Goffman et al., 1978].

[Hogan, 2010] extends Goffman’s ideas to social media, noting that “actors” now create “artifacts”: persistent records which are accessible to others. Hogan transforms Goffman’s metaphor to make individuals creators of “exhibitions” that are available online rather than live “actors” on a stage. One key difference between these approaches is that Goffman’s model has individuals acting to control perceptions of those directly in front of them whereas Hogan notes that online profiles must span time and audience. Actors can take breaks between shows and put on a new performance for a new audience - in Goffman’s model individuals are similarly able to keep impressions held by different audiences somewhat separate. For Hogan, online profiles are more similar to recordings of shows. They are more permanent, often fail to capture the nuance of the original experience, perceptions of them may change over time, and they may be replicated and shared with others. In many cases, the full audience of an online post may never be known and for this reason users tend towards pleasing the “lowest common denominator” [Hogan, 2010]. If a user wants to make a political post on Facebook that they think their friends will enjoy, they also must consider that their co-workers, family, classmates, and exes are able to see this information. Further, any of these individuals may show the post to their network and the post is available to all *future* connections the poster may make. With this in mind, perhaps the user will choose to share an innocuous picture of their pet instead. Hogan suggests that users act based on mental models of the site they are posting to, considering both how the site curates their posts and a few “specific salient individuals” who are likely to see their post [Hogan, 2010].²² Different sites may have different audiences and

²²These individuals may represent larger groups. Rather than considering the reaction of one’s entire extended family, a user may ask simply “what would my mother think?” before posting.

norms - political activists on Twitter may be quiet on Facebook if some salient individuals are on one platform but not the other - but the underlying principles are the same.

[Cheney-Lippold, 2011] explores the relationship between cultural identity and algorithmic identity. Algorithms automatically construct user profiles across certain attributes (especially age, gender, and income) and assign users to different categories of being (i.e. “male” or “teenager”). These profiles estimate cultural “ground truth” categories of being, traditionally defined by cultural norms and individual self-identification. Observing that “code can also construct meaning,” Cheney-Lippold argues that creating these profiles shapes what it means to belong to different categories online. If automatically generated profiles change the advertisements we see, they help to shape our sense of self and enforce normative standards for belonging to a certain category. More than that, they create a feedback loop wherein definitions of categories “can shift according to the logic of the algorithm” [Cheney-Lippold, 2011]. Algorithmic profiling necessarily creates in-groups and out-groups and treats them differently. This is problematic in part because profiles do not necessarily reflect the self-identification of the person being profiled, may be constructed without their awareness, and offer little opportunity for recourse. “We are effectively losing control in defining who we are online, or more specifically we are losing ownership over the meaning of the categories that constitute our identities” [Cheney-Lippold, 2011]. Cheney-Lippold is not entirely bleak and insists that the definitions of categories change as new information is added. He appears to observe a sort of “back and forth” between cultural norms and information used to generate algorithmic profiles. Advertisers want profiles to be accurate to potentially changing ground truth identities, but those identities are in turn influenced by online interactions. With this in mind, it is not obvious how often profiling techniques reconsider key characteristics such as gender. On a platform such as Facebook, does this profile rely entirely on user-specified data, or is it also based on the content of posts or information from data brokers? Each case has different implications for how often, if ever, the profile is updated. This lack of clarity makes it difficult to determine how much social media profiles are altered by cultural changes.

Together, these three works highlight the importance and difficulty of controlling social media perception. If we accept that managing impressions is as important today in Western culture as Goffman proposes, then it is clear that users must take care to manage their social media profiles. Both Hogan and Goffman suggest that individuals attempt to manage impressions based on an understanding of how their actions will be perceived by others and how that perception will influence future actions. However, in the case of algorithmic profiling, it is not reasonable to expect that many users will have good mental models for how algorithms generate impressions (if users realize such impressions are being generated at all). Even if users are aware of the importance of

these profiles, there are few resources to allow them to build a better understanding of how profiles are created. This project aims to equip users with tools to better manage these impressions.

Chapter 3

Methodology - Prototype

This chapter outlines the prototype tool, which is called Mockingbird as it helps Twitter users mimic the sounds (tweets) of other types of users. First I outline the design of the prototype, explaining each algorithmic profile employed, the explanations offered, and the tools available for editing tweets. Then I explain how different parts were implemented, including how some classifiers were trained.

3.1 Prototype Design

Mockingbird is designed as a free-standing website for ease of use and accessibility. Mockingbird was built using Django¹ and is hosted at <http://mockingbird.hip.cat>. The Python script which works as the server is on a Linux virtual machine.²

This tool aims to allow users to focus on the algorithmically generated profiles they find most interesting and so is set up to increase in detail as the user engages with a particular profile. For instance, after the user confirms which tweets are theirs, they are presented with all available profiles grouped by attribute. They can then click on each profile for more information and have the option of clicking again to see an explanation or attempt to alter the profile. This way users see only as much information as they desire and are not forced to spend time looking at profiles that do not interest them. See Section 4.2 for more detail, including screenshots.

Other than the underlying website, Mockingbird relies on two types of tools: algorithmically generated profiles and explanatory tools. Each is outlined below.

¹Django was chosen because the available models were implemented in Python and Django backend is in Python. For more details on Django, see <https://djangoproject.com>.

²All relevant files are provided in the directory `mockingbird` included in the code for this project.

3.1.1 Generating Algorithmic Profiles

As outlined in Section 2.2, there are a wide range of ways to profile on social media data and an even wider range of topics on which to profile. For this experiment, a limited number of characteristics had to be chosen for several reasons. Firstly, when conducting the lab experiment, only a fixed amount of time is available and it may be easy to overwhelm participants with too many profiles. As this experiment is meant to also allow users to experiment with changing their algorithmically generated profiles, it would be counter-productive to spend too much time looking at unimportant or uninteresting profiles. Secondly, as this research does not aim to introduce any novel tools or datasets for generating social media profiles, the available profiles are limited to those which have been previously studied. Gathering and labelling data for use in this experiment would be too expensive in terms of time and resources. Furthermore, this research is meant to give users tools to change and understand existing profiling tools. While the details of the profiles used in practice are largely kept secret by the companies employing them, it is logical to assume that they are similar to existing tools available online or outlined in the literature. Finally, care must be taken concerning profiling users based on sensitive characteristics. Although corporations profile based on special category characteristics such as political leaning and race [Dewey, 2016], finding datasets and building classifiers on them raises ethical concerns. This is especially true of characteristics which users may wish to hide on social media, including sexual orientation and drug abuse (profiled in [Kosinski et al., 2013a]). Such classifiers, once trained, could be misused in the future for discriminatory or otherwise unethical purposes. Even if the classifiers themselves are destroyed after the experiment, using them to profile people in a closed setting could cause distress on the part of the profiled individuals. While some amount of this may serve to reinforce the importance of controlling user profiles, this point can also be achieved with less invasive techniques. For this reason, extremely sensitive topics were avoided and some potentially uncomfortable profiles were included only with explicit user consent (see Section 4.1.2).

The following profiles were chosen due to ease of implementation and the perceived value of the attributes they profiled. Some consideration was also given to including diverse methods of profile generation. The classifiers fall into a few broad categories, explained in detail below.

Lexical Classifiers

Lexical classifiers, as introduced in Section 2.2.2, are one of the main approaches used by this project. The gender and age lexica provided in [Sap et al., 2014] were used directly.³ This project

³This project uses a slightly different tokenizer than in [Sap et al., 2014], however this is not expected to have a significant effect on performance.

also makes use of a sentiment lexicon generated by the National Research Council of Canada and trained on Twitter data [Kiritchenko et al., 2014, Mohammad et al., 2013, Zhu et al., 2014]. This lexicon, called NRC Emoticon Lexicon or Sentiment140 Lexicon, was also generated using an SVM classifier and makes use of both 1-grams and 2-grams. Approximately 1.6 million tweets were used to generate this lexicon, with each tweet being identified as positive or negative using emoticons as seeds [Kiritchenko et al., 2014]. The inclusion of bigrams is particularly important for sentiment classification, as often negations change the meaning of a phrase (such as “not bad” being neutral, if not positive, when compared to “bad”). To incorporate both bigrams and unigrams, the scores of the unigram lexicon and the bigram lexicon are computed separately and summed to get a final classification.

Additionally, this project uses the sentiment and objectivity analysis tools provided by the `pattern` Python library [Smedt and Daelemans, 2012] through the `TextBlob` Python library⁴. This library utilizes a lexicon to analyze adjectives and returns a sentiment score between -1 and 1 and a subjectivity score between 0 and 1. A score of -1 means the text is extremely negative whereas 1 means extremely positive. At a subjectivity score of 0, the text is entirely objective; at 1 it is entirely subjective.

In general, lexicons have the advantage of being both easy to compute and perfectly interpretable. Simply looking at the words in the text which appear in the lexicon and their weights provides an entirely interpretable explanation. However, lexicons are rigid in that they only work if the text to be classified has some of the words in the pre-trained lexicon. Lexicons are unable to pick up on new slang, misspelled words, or typos. This makes them vulnerable to simple attacks that break tokenization [Hosseini et al., 2017]. As such, they are useful for baseline models but are unlikely to be robust to outliers or attempts at deception.

IBM Classifiers

A suite of text classification tools called Personality Insights⁵ is provided as a free API by IBM. This API classifies an arbitrary text sample of at least 200 words on several personality characteristics, including the Big Five [McCrae and John, 1992], needs, and values. For the purposes of this experiment, only Big Five results are shown due to their relevance to Cambridge Analytica and other ad profiling systems [Cadwalladr and Graham-Harrison, 2018]. To generate input for the API, all tweets are concatenated in to one string with each tweet separated by a newline character. This is then passed as input to the API, which returns a JSON object with percentiles for the

⁴<https://textblob.readthedocs.io/en/dev/>

⁵<https://personality-insights-demo.ng.bluemix.net/>

Big Five characteristics. These percentiles are presented directly to the user, along with a brief description of each characteristic and its implications for user behavior.

Although the official documentation for the Personality Insights API is lacking in details⁶ it is likely that the model is based on [Arnoux et al., 2017], as covered in Section 2.2.2.⁷ This work uses GPs for Big Five personality predictions with an average correlation of 0.33. Despite the limitations of the IBM model, it provides insight into a relevant area of profiling. Personality models are important not only because they are being used to target individuals but also because they may be seen as especially private or invasive. [Warshaw et al., 2015] describe personality models as the first steps towards “hyper-personal analytics systems” and found that half of study participants felt uncomfortable with the accuracy of their personality profiles. The inclusion of Big Five profiles in this experiment allows users to engage directly with potentially uncomfortable or intimate algorithmically generated profiles.

Machine Learning Classifiers

In recent years, the usage of machine learning in generating algorithmic profiles has become ubiquitous. All told, seven machine learning classifiers were trained for this project in addition to one provided by the TextBlob library. This eighth classifier trains an NBC on the NLTK movie reviews dataset⁸ and returns the probability of a given text having negative or positive sentiment. This package was used directly to generate the sentiment probabilities, with the model being trained once and evaluated separately on each tweet and collection of tweets.

The main challenge of training machine learning classifiers is the lack of available data. Datasets are expensive to create in terms of time and resources, and few existing Twitter datasets are available due to privacy concerns and changes to Twitter’s terms of service. The decision to use existing datasets restricted profiles to those based on previous work that provide labelled data. This fits with the goal of the project to reflect existing methods of user profiling rather than creating or explaining new ones.

The seven classifiers are neural networks trained on income data from [Preoȕiuc-Pietro et al., 2015b]. These data were obtained by combining the files provided by [Preoȕiuc-Pietro et al., 2015b]⁹ with files pointed to by [Preoȕiuc-Pietro et al., 2015a].¹⁰ The first of these contains user_ids along with predicted demographic information. The second contains bag-of-words representations of the tweets

⁶The documentation says classification uses a “machine-learning algorithm,” see <https://cloud.ibm.com/docs/services/personality-insights?topic=personality-insights-scienceresearchInfer>.

⁷This assumption is based on the fact that the documentation cites this paper and that the paper provides a link to their API which points to the IBM Cloud Services homepage.

⁸See https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.en.sentiments.NaiveBayesAnalyzer for more information.

⁹From https://figshare.com/articles/Twitter_User_Income_Dataset/1515997

¹⁰From <http://www.sas.upenn.edu/~danielpr/jobs.tar.gz>

for each user along with a dictionary to convert from bag-of-words tokens to text. This choice of data has several limitations. First, because only bag-of-words representations are available, it is not possible to leverage machine learning techniques such as Recurrent Neural Networks (RNNs) that make use of sequential data. Information was irretrievably lost when the tweets were stored this way and so a full analysis is not possible. Tweets may have been stored this way to increase the difficulty of linking them to Twitter users. Because tweets are public, it is almost always possible to “de-anonymize” users based on a tweet by simply searching for the text of the tweet. Reconstructing a single tweet from a bag-of-words of hundreds may take considerable effort. Even with this limitation, bag-of-words data allows many possible classifiers, including feed-forward neural networks. In addition to losing word order, it is not possible to tell how many tweets the bag-of-words data is meant to represent for each user. For two `user_ids`, no bag-of-words data was provided and the number of words varies across users. This uncertainty was mitigated by using a binary term frequency term (see Section 3.2.1).

A second limitation is the nature of the data labels. Income data, represented as an integer, was collected by gathering Twitter profiles with occupational keywords in their bios. After hand-pruning these profiles, each was assigned the mean income of their occupation according to UK job data [Preoțiuc-Pietro et al., 2015b]. As such, estimated incomes do not take into account factors such as the location or experience of individual users. They may also be biased towards occupations that use Twitter frequently or who are likely to include their occupation in their bio. Even with these considerations in mind, the income labels seem to be a reasonable estimate of user income. The other labels are less sound. For several attributes, a probability distribution is given across classes. These probabilities are meant to be used as features for income regression and come from pre-trained log-linear classifiers developed in [Volkova et al., 2015]. These classifiers were trained on 200 tweets from 5000 users which were labelled using crowdsourcing. Their performance is not reported in [Volkova et al., 2015] and the models are not available directly.¹¹

The first problem with using these labels is that the income dataset is not balanced in terms of classes. For instance, of the nearly 5200 users only two were labelled as having children. Less extreme class imbalances were managed with class weighting and combining classes. For race data, the overwhelming majority of users (5112/5189) were classified as “white” rather than “asian,” “black,” “hispanic,” and “race_other.” In some cases, I addressed this imbalance by combining classes to reduce the problem to binary classification. All profiles that had a probability of being “white” greater than or equal to 0.5 were assigned the label “white” and all others were assigned

¹¹On the primary author’s web page, it states that the models are available upon request but she did not respond to an email request.

the label “non-white.” This resulted in 1769 “non-white” examples and 3420 “white” examples. While still unbalanced, this split makes it possible to train a non-trivial classifier. This solution was not viable for all attributes and so some were necessarily omitted. See Section 3.2.1 and Tables 3.4 and 3.5 for more details on handling class imbalance. The second problem is that any classifier learning from this data is not learning from “true” labels. Instead, it is learning to reproduce the labels classifiers from [Volkova et al., 2015] produced on data not used for training and without ground truth labels. It is not possible to identify which labels in the data accurately reflect the ground truth of the user they are profiling. Any errors in the original classifiers are likely to be transmitted to classifiers trained on this dataset.

Despite these problems, the income dataset was deemed too valuable to omit. Social media data is often messy and errors in profiling are unavoidable. It is reasonable to expect that real-world models are trained with noisy data, especially if those data are inferred from user information or third-party data brokers. Existing social media profiles frequently contain errors [Rao et al., 2015] and companies do not release information on how accurate their classifiers are. The purpose of this experiment is to see how users respond to algorithmically generated profiles and it is important to see their reactions to both accurate and inaccurate profiles. The labels offered by this dataset include income, education level, relationship status, race, religion, and political affiliation. At least three of these (race, religion, political affiliation) are considered special category data and required special handling (see Section 4.1.2). Although each of these categories are frequently used for ad profiling [Dewey, 2016] no other datasets were available due to their sensitive nature. For information on how these classifiers were trained, see Section 3.2.1.

M3 Classifiers

While the other methods are based on the social media text, in this case raw tweets, the M3 classifier uses the profile picture, biography, username, and screen name to profile users. See Section 2.2.3 for full details on how this classifier works. Each category offered by the M3 network is provided to Mockingbird users. This project uses directly the pre-trained models and API provided on GitHub.¹²

The M3 network is especially valuable for this experiment because it does not rely on tweet data and so shows users an example of how other features may be used for classification. Comments on the M3 results may be particularly interesting, as the data used is likely to already be well-curated by users to craft their online identities. Profile pictures are likely to be carefully selected images of the user, and the biographies likely to contain personal information. A username may be difficult

¹²<https://github.com/euagendas/m3inference>

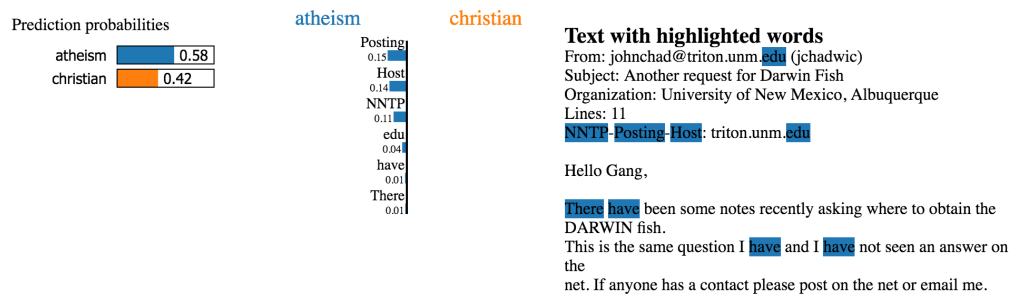


Figure 3.1: An example of LIME visualization, from the documentation.

to change and can be used to link one user across various online accounts. These characteristics (except for the biography) are visible every time a user tweets and therefore are the most obvious and consistent representation of their online persona. They are also what is primarily used to connect the real-world lives of individuals to their online personas. Without a picture, a biography containing personal details, or an obvious username, it is difficult to establish a definite link between an offline connection and an online account. For these two reasons, one would expect users to be less willing to change these attributes of their account. While individual tweets often represent a temporary situation or thought, profile information is a persistent proxy for personal attributes.

3.1.2 Explanatory Tools

Although simply showing profiles has a degree of educational power, without explanations they offer little opportunity to aid user autonomy. The primary goals of providing explanations as part of this experiment are to help users change their algorithmic profiles and to allow them to make better judgments about the trustworthiness of algorithmic profiles.

Explanations for lexical classifiers are quite straightforward. To explain why an instance was labelled negatively, for example, one needs only to return the words with the highest negative weight. These explanations can be generated in real time or afterwards. To get words most representative of each class requires only a lookup in the lexicon of the words with the most extreme weights. When considering weights, the program returns the terms with the highest weights adjusted for frequency. This occasionally results in the most explanatory word being a common English stop word such as “of”. In this way, explanations of profiles tend to focus on broad trends rather than specific, infrequent words. Infrequent but heavily weighted words appear when users look at explanations for individual tweets.

Explanations for machine learning classifiers were generated using the LIME technique (see Section 2.1.3). Explanations were offered with twenty-five words at the profile level and with

five words at the individual tweet level - these limits were chosen to keep visualizations brief and manageable. Explanations, as well as visualizations for them, were generated directly from the LIME Python package.¹³ See Figure 3.1 for an example of this visualization.

There were a few challenges to using the LIME library directly for explanations. The LIME technique is by nature computationally expensive, especially on images. Due to time constraints imposed by lab work, all available explanations were created on demand so that time would not be spent unnecessarily generating LIME explanations. Even with this optimization, there was not enough time to provide explanations for the M3 profiles. Allowing explanations for M3 profiles would have required allowing users to change their profile data, which would have necessitated substantial additions to the system underlying the web page. It is also not clear that LIME could make meaningful sense of character-level embeddings with its standard text explainer. This would be crucial when the entire input for username is a single word - if LIME explains at the word level, no useful information could be gleaned from changes to the username.

Additionally, no explanation was made available for the IBM profiles. This is because the IBM API is limited to 1000 requests per month. The LIME technique would have to be run separately on each of the five attributes, and each of these attributes would likely generate a few hundred to a few thousand perturbations. There was no way to make explanations available to all users with this limitation. This was also deemed acceptable because the IBM API requires at least 200 words to build an accurate profile, preventing classification and explanations for individual tweets. This would have only limited usefulness in helping users to control their Big Five profiles.

The LIME package is not equipped to handle regression on text input. To circumvent this problem for the income regression model, I cast the integer valued regression output into two categories of “above average” and “below average” and assigned confidences based on the predicted value. For instance, if the predicted value is exactly the average, each class has 50% confidence. If the predicted value is 0 pounds, the confidence of low income is 100% and if the predicted value is the maximum possible, above average is predicted with 100% confidence. Intermediate values are scaled accordingly - a predicted income of half of the average would be classified as low income with confidence 75%. This is meant to serve as a rough estimate for LIME to use rather than as a replacement for the integer value provided by the regression.

When users opt to generate LIME explanations in the lab session, I made use of the wait time by asking questions about the experiment so far or about what they expected the generated explanations to say.

¹³<https://github.com/marcotcr/lime>

3.1.3 Editing Tools

Although explanations may help users understand algorithmically generated profiles, they do not always provide a clear way to change these profiles without removing all relevant words. To help with this for lexical classifiers, a general suggestion to try misspellings or adding spaces and periods was included at the top of the editing page, inspired by [Hosseini et al., 2017]. Apart from these general tips Mockingbird provides two main tools to help users gain control over their algorithmic profiles: synonym suggestion and style translation, both offered for lexical classifiers.

Synonym Suggestion

For each tweet, synonyms were suggested for the most explanatory words when available. To generate synonyms, the program first gets a list of all synonyms for these words using WordNet [Miller, 1995]. WordNet is a “lexical database” which clusters words based on their meanings and maintains hierarchical relations between related words. To generate synonyms for an explanatory word, the program looks up that word in WordNet and collects all of its “synsets.” Each synset represents a set of synonyms for a given meaning of the explanatory word; taken together, they represent all known synonyms for the explanatory word in WordNet [Miller, 1995].

The program then calculates a score s for all synonyms that are in the lexicon and contribute toward the target class. In the binary classification case where a negative value corresponds to a label of 0 and a positive value corresponds to a label of 1, a synonym t is only considered if $\text{sgn}(l(t)) = \text{sgn}(y_t)$ where $l(t)$ represents the value of the lexicon for synonym t and y_t is the target class. The score s is calculated according to the following equation:

$$s = \begin{cases} |l(t)| * \text{count}(t), & \text{if } \text{sgn}(l(t)) = \text{sgn}(y_t) \\ 0 & \text{otherwise} \end{cases}$$

where $\text{count}(t)$ is the number of times t is suggested by WordNet. This score was chosen because often WordNet suggested unexpected synonyms from an uncommon or archaic use of the explanatory word. While these unhelpful synonyms are difficult to remove entirely, WordNet also returns the same synonym multiple times if it is a synonym for several definitions of the explanatory word. Following the assumption that words that show up more frequently as synonyms are more likely to fit the sense of the word being replaced, these frequented words are up-weighted. The top five words with the highest non-zero score, regardless of which word they are meant to replace, are displayed to the user. Each synonym suggestion is shown as the original word, a right arrow, and then synonym so that it is clear which word the synonym is meant to replace. Due to this method

for choosing synonyms, all suggestions will be sure to contribute towards the target class. See Section 3.3.1 for information on how this approach worked in practice.

Style Translation

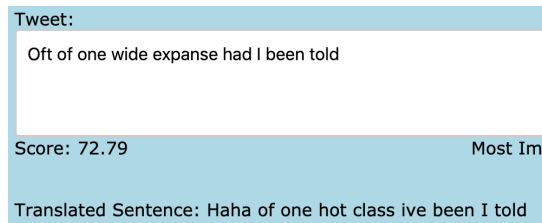


Figure 3.2: An example translation from old to young.

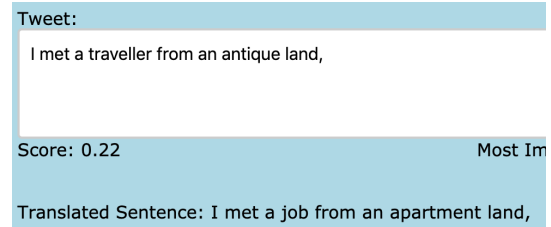


Figure 3.3: An example translation from young to old.

The other major editing tool is an automated translation between “male” and “female” and between “young” and “old.” [Shetty et al., 2018] make available pre-trained models for gender and age on their GitHub,¹⁴ although the code required many fixes to be usable for this project. The provided code is in Python 2.7 and has no documentation - it was necessary to convert this code to Python 3. The models will only load on a GPU, so it was necessary to load them in Google Colab and then save them as CPU models. The adapted code can be found included with this thesis in the directory `classifiers/A4NT`. These models are trained on a blog dataset of approximately 14,000 blog posts and are designed to change a sentence from one class to another while maintaining semantic validity (see Section 2.3 for more details on how this model works and Section 3.3.1 for its shortcomings).

3.2 Prototype Implementation

To implement the design outlined above, several classifiers required training, the details of which are included below. Additionally, in the course of building Mockingbird and testing with a “test” profile a few problems with editing tools were discovered.¹⁵

3.2.1 Training Neural Networks

All models are standard feed-forward neural networks with either one or two hidden layers with ReLU activation functions. Dropout was performed after each hidden layer to aid regularization.

¹⁴<https://github.com/rakshithShetty/A4NT-author-masking>

¹⁵All examples shown here and elsewhere are from a test account with fake tweets. Each test tweet is a line from a poem or literary quote. These are meant to illustrate how the system works without revealing any personal information.

Other than income, all are categorical results. Income was handled both as a regression problem and as categorical classification problem with three classes (“below average,” “above average,” “far above average”). Each categorical classifier has a final layer with a softmax activation function. For the regression, training data was scaled to be $[0, 1]$ and a sigmoid activation function was used on the output layer. This output was then converted to a predicted income by multiplying it by the maximum income in the dataset (£111,413). All models were trained for twenty epochs with categorical cross entropy loss and the Adam optimizer [Kingma and Ba, 2014] with default hyperparameter values. All models were implemented in `keras` [Chollet et al., 2015] and run on Google Colab with a GPU hardware accelerator. See the included file `NN.ipynb` for an example.

Each model used a TF-IDF representation of the bag-of-words with English stop words removed. TF-IDF stands for term frequency - inverse document frequency and gives higher weight to words in a document¹⁶ which are used infrequently in other documents. Term frequency can either be binary (an indicator of whether or not this term appeared in the document) or a count of the number of appearances. This TF-IDF representation was obtained by converting each bag-of-words to text by looking up each index in the index-to-word dictionary and repeating the entry the appropriate number of times.¹⁷ These text entries were then used to train a `sk-learn` TF-IDF Vectorizer which was used to convert all text to its TF-IDF representation. Binary term frequency was used throughout as it was seen to generally boost classifier performance.

All models train on 90% of the data, with 10% of this set aside for validation. After tuning hyperparameters on the validation data, the models are trained on the full 90% of the data and evaluated on the remaining 10%.

To compensate for class imbalance, class weights were used. Focal loss [Lin et al., 2017] was also implemented but did not seem to improve performance and so was not used for the final training. Models generally fit well on the training set but had trouble generalizing. In addition to dropout, kernel regularization was attempted but it tended to substantially degrade performance. The relatively poor generalization is likely due to the small number of examples, the sparsity of bag-of-words features, and the noisiness of the data labels.

Each model was tested over the hyperparameters displayed in Table 3.1, except for the regression model which was also tested with 1024 hidden units.

¹⁶Here a document is a concatenation of the text of all tweets by a given user.

¹⁷The resulting text was equivalent to “sorting” the original text by the index correspond to each word. The fact that the text was out of order compared to the original tweets has no impact on the TF-IDF weighting.

Hyperparameter Name	Tested Values
Dropout	0.0, 0.2, 0.3, 0.4, 0.5
Batch size	32, 64, 128
Units	64, 128, 256, 512
Hidden Layers	1, 2

Table 3.1: Hyperparameters tested for each neural network.

For all models with two hidden layers, both layers have the same number of units and have a dropout layer after them. The optimal hyperparameters can be found in Table 3.2. Table 3.3 reports accuracy and F1 score (the harmonic mean of precision and recall) for both the validation and the test set. When the hyperparameters that resulted in the highest accuracy did not produce the highest recall, parameters that best balanced the two metrics were chosen. F1 score is not available for the categorical income data. It is clear that both accuracy and F1 vary substantially across attributes.

Name of Model	Hidden Layer Count	Hidden Units	Dropout Rate	Batch Size
Income (Regression)	2	1024	0.2	32
Income (Categorical)	2	256	0.0	64
Education	2	512	0.4	128
Relationship Status	1	128	0.2	64
Political Affiliation	1	128	0.0	32
Race	2	64	0.2	128
Religion	2	512	0.2	64

Table 3.2: Tuned Hyperparameters.

Name of Model	Val Accuracy	Val F1	Test Accuracy	Test F1
Income (Categorical)	71.68%	-	73.41%	-
Education	62.10%	73.86	58.96%	70.77
Relationship Status	75.80%	25.17	75.53%	24.10
Political Affiliation	55.67%	61.16	52.60%	53.21
Race	60.17%	34.04	56.84%	32.26
Religion	55.03%	59.92	58.19%	55.71

Table 3.3: Test Accuracy and F1.

Due to the class imbalance, care was taken to ensure that the classifiers are not trivial (only returned the most frequent class). Trivial classifiers, if more accurate, would not be as interesting to users as it would not be possible for users to change inaccurate classifications. Class labels were determined by selecting the class with the highest probability. This often resulted in classes labels taken from a plurality but not a majority. Several classes were combined to reduce imbalance as outlined in Section 3.1.1. This was achieved by summing the probabilities of all of the classes to be combined and making that the probability of the new class. Combinations were chosen to make sense with the meanings of the labels. For instance, the classes “undergraduate education”

Attribute Name		Original Classes				
Religion	<u>Christian</u>	<u>Hindu</u>	<u>Jewish</u>	<u>Muslim</u>	<u>Other</u>	<u>No Religion</u>
	2600	0	0	0	0	2589
Race	<u>Asian</u>	<u>Black</u>	<u>Hispanic</u>	<u>Indian</u>	<u>Other</u>	<u>White</u>
	0	77	0	0	0	5112
Relationship	<u>Single</u>	<u>Divorced</u>	<u>Dating</u>	<u>Married</u>	<u>Other</u>	-
	5163	0	7	2	17	-
Politics	<u>Conservative</u>	<u>Independent</u>	<u>Liberal</u>	<u>No Political</u>	-	-
	819	0	0	4370	-	-
Education	<u>High School</u>	<u>Undergraduate</u>	<u>Graduate</u>	-	-	-
	4535	654	0	-	-	-

Table 3.4: Class breakdown for original (imbalanced) classes.

and “graduate education” were combined to “college.” These are reported in Table 3.4. The final distributions along with the class weights are included in Table 3.5. Below information for the income models is included separately, as these classifiers had to be handled slightly differently.

3.3 Income

The income data in pounds sterling (GBP) was reported as an integer between 8,395 and 111,413 with an average value of 32,517 and a median value of 28,959. Two neural networks were trained on this data. The first was a regression model for which all income labels were divided by the maximum to scale the results to $[0, 1]$. This model achieved an average absolute error of £9,707 and a median error of £5,984. This is comparable to the mean average error reported by [Preoțiuc-Pietro et al., 2015b].

The second income model was categorical. All users with an income less than the average were assigned label 0, users with an income between average and £75,000 label 1, and users earning above £75,000 label 2. Labels were converted to a one-hot encoding for classification. There were 3,264 below average examples, 1776 above average examples, and 149 far above average examples in the dataset. The class weights were set to 0.8 for below average, 1.1 for above average, and 1.5 for far above average.

3.3.1 Analysis of Editing Tools

Overall, editing tools performed relatively well. However, both had some shortcomings worthy of comment.

Attribute Name		Final Classes	
Religion	count	<u>Non-Christian</u>	<u>Christian</u>
	weight	2600	2589
Race	count	<u>Non-White</u>	<u>White</u>
	weight	1	1
Relationship	count	<u>Available</u>	<u>Taken</u>
	weight	1769	3420
Politics	count	<u>Non-Political</u>	<u>Political</u>
	weight	0.75	0.25
Education	count	<u>High School</u>	<u>College</u>
	weight	3190	1999
	weight	0.75	5

Table 3.5: Class breakdown and weights for final classes.

Synonym Generation

Suggested Synonyms: flash→dash, like→wish, flash→heartbeat, like→same, flash→twinkle

Figure 3.4: An example of sensible synonyms

Suggested Synonyms: birds→skirt, birds→raspberry, birds→bird, birds→chick, birds→boo

Figure 3.5: An example of unhelpful synonyms

Suggested Synonyms: like→wish, or→surgery, by→aside, like→same, by→past

Figure 3.6: An example of confusing synonyms

In practice, the results for synonyms are mixed. Mostly they are somewhat sensible as in Figure 3.4. They don't necessarily agree in part of speech, but the meaning is comparable and it is not hard to imagine using the suggested word or a variation of it as a substitute. In some cases, they are unhelpful because the word to be replaced has a large number of meanings as in Figure 3.5. "Bird" has many uses in English, including some slang ones, and it is unlikely that the usage intended in the tweet will be represented in the synonym list. Note also that some of these synonyms are outdated or misogynistic - it is up to the user to determine which available synonym best suits their

needs. Occasionally synonyms are confusing because they are not domain specific or do not make sense. In Figure 3.6, WordNet shows “surgery” as a synonym for “or.” It is likely that WordNet interpreted “or” as O.R., an abbreviation for “operating room,” and so offered “surgery” as a synonym. This chain of logic is not presented to the user, and the suggestion is likely confusing unless the user has domain specific knowledge. In another example, the method suggested replacing “x,” used as slang for “kisses” at the end of a message, with its value as a Roman numeral. These problems could be overcome with a more complex model, for instance one that leverages semantic loss suggested in [Shetty et al., 2018] (see Section 2.3) or makes use of a different tool to suggest synonyms. This is left to future work, as synonym generation is a secondary element of this project.

Style Translation

Style translations were found to suffer from several main problems.

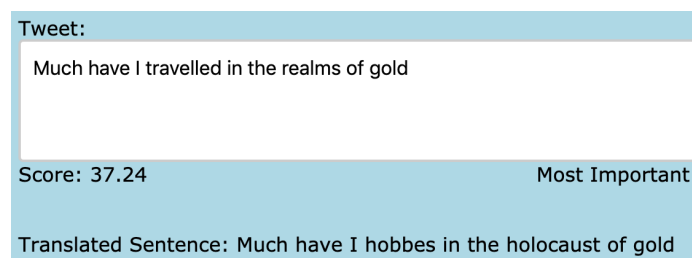


Figure 3.7: An example of a possibly offensive translation.

First, the blog dataset is not ideal for this application. It was collected almost fifteen years ago and this often results in outdated translations. It also contains words that are generally considered offensive today, which have been learned by the translator. This is particularly prevalent when translating from “old” to “young,” indicating that young blog users may have used offensive or “edgy” words (see Figure 3.7). This occasionally results in translations with offensive words that were not present in the original text. Although this issue reflects the reality of the dataset, and to a lesser extent Internet culture in general, automated translation with such serious problems is likely to impact other filters, such as toxic speech detection, and may result in a social media profile much worse than the original. It is not hard to imagine cases in which an automatic translator introducing an offensive word could severely damage the reputation of its user.

Second, the translator was trained on individual sentences which are shorter than typical tweets, resulting in truncated translations. I attempted to fix this problem by dividing long tweets into shorter sentence fragments by splitting on characters such as commas and colons. Although this improved performance on longer tweets, the translator has a tendency to replace unknown words

with periods. This makes translated sentences look broken and often destroys their semantic validity. These problems can be more or less important depending on the style of the Twitter user - the translator is likely to work better on short tweets with simple grammar and common words.

Third, the models have a limited vocabulary and require replacing certain text features with generic tags. The vocabulary is problematic due to the age of the dataset it was trained on and the frequency of slang and hashtags on Twitter. Generic tags are a problem when numbers are changed from their decimal representation to the tag “NUM,” for instance. The generated sentence may have several of these “NUM” tags, and possibly more or fewer than the original sentence. To handle this in the case of numbers, emojis, URLs, hashtags, and user tags, I captured each of these special characters in the original tweet and substituted these in for the replacement tags.

Consider the following example to illustrate this. If the original tweet is “I’m taking the 10:15 train @friend pic.twitter/123” it would be preprocessed into the following: “I’m taking the NUM:NUM train @friend TWIT_PIC”. This preprocessing, replacing some characters or words with a generic tag, is necessary for the pretrained translators. The translated sentence may be “I’m taking the NUM:NUM NUM train @ TWIT_PIC”. The program would then search the original tweet using regular expressions to replace each generic tag in the translated sentence with an appropriate word or character from the original sentence. After observing that user tags and hashtags were always translated to just “#” and “@”, these were also identified with regular expressions and replaced. Here, the translation returned to the user would be: “I’m taking the 10:15 NUM train @friend pic.twitter/123”. The first two “NUM” tags were replaced with the two numbers in original tweet but as the third tag did not have a corresponding number it was left alone. For emojis, URLs, hashtags (beginning with #), and user tags (beginning with @), this approach was effective as the translated sentences rarely contained more instances of the tag than the original sentence. This was not the case for “NUM” as several translated sentences introduced new “NUM” tags (see Figure 3.10).

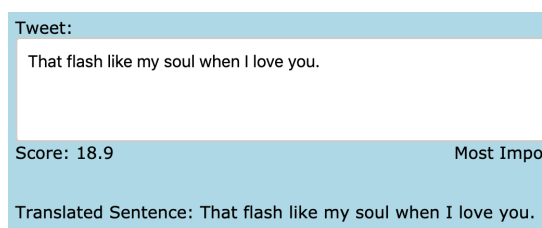


Figure 3.8: An example translation with no changes.

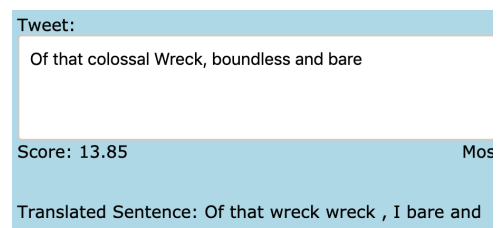


Figure 3.9: An example translation with repeated words.

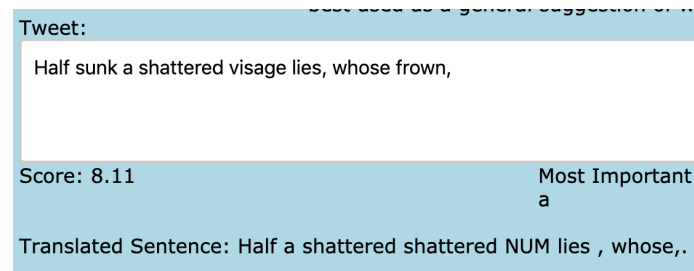


Figure 3.10: An example translation with a “NUM” tag.

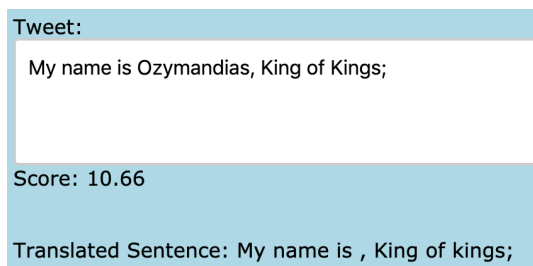


Figure 3.11: An example translation that omitted a new word.

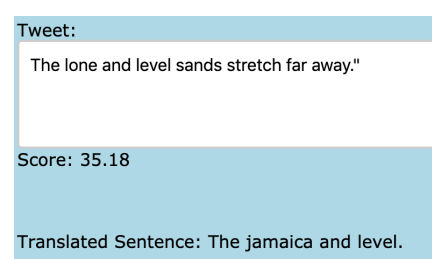


Figure 3.12: An example translation which makes no sense.

There are a few other, broader problems with translations. In some cases, the translator recreates the original tweet perfectly (Figure 3.8). This may indicate an ambiguous tweet for the A⁴NT network but is not useful to anyone attempting to change their profile. Words that do not occur in the A⁴NT vocabulary are often entirely omitted by the translation (Figure 3.11). Users may be unable to differentiate between situations where the translator removes words to change the classification and situations where unknown words are left out. The translations occasionally suffer from stuttering (Figure 3.9) and in the worst case are completely incoherent (Figure 3.12). Even when translations are somewhat reasonable (Figures 3.2 & 3.3), they deviate substantially from the original meaning of the translated sentence.

For all of these reasons, the translations generated by these models were mediocre at best and, at worst, almost unrelated to the original sentence. They were still included for several reasons. Even if the translated sentence was incoherent, it still may provide hints as to how the original sentence can be changed - for instance, word substitutions or deletions. They also give cues as to what sort of words are appropriate for the target class - Figure 3.3 indicates that words like “job” and “apartment” are associated with older users, whereas Figure 3.2 suggests “haha,” “hot,” and “class” as younger words. Since users are ultimately in control, they can decide how much or little thought to give to the translation. Translations also help to inform users about the state of automated style text translation. Seeing insensitive or faulty translations makes clear the limitations and risks of using such methods. The experimental risk is that subjects become more skeptical of how

technology used in this experiment compares to the state-of-the-art. If subjects expect Google Translate-level translations and see these instead, they may think that all of the classifiers used in this experiment are similarly under-performing relative to those in industry. However, I believe this concern is outweighed by the possible benefit of providing these translations.

Chapter 4

Methodology - Experiment

This chapter provides an overview of the lab experiment for testing the Mockingbird tool. The first section outlines the details of the lab work including recruitment and handling of special category data. The second section provides a brief look at the Mockingbird prototype, including numerous screenshots, so that the reader will have an idea of the interface and user experience.

4.1 Experimental Design

In order to address the research questions using the outlined profiling tools, I conducted a lab study of a handful of participants ($n = 6$). A lab study was chosen for several reasons. First, some methods used in this experiment are rather slow, especially LIME explainers. This would make conducting a large survey online difficult, as users may have to wait for several minutes without the website responding. In lab, this pause provides an opportunity to ask questions and gauge the user experience so far. Second, managing online profiles is a personal and somewhat subjective exercise. This project aims to see not only what decisions people make, but also why they make them. The lab experience allowed me to ask follow-up questions to better understand the preferences and reasoning of users. This conversational style also facilitates open-ended user feedback into what improvements and changes they would like to see and allows them to ask questions about the methods being used. Conducting the experiment in the lab affords better control with regards to sensitive data and classifiers over sensitive attributes. All tweets are easily re-identifiable, so if a single full tweet were released a participant could be de-anonymized. More importantly, three of the profiles (political affiliation, race, and religion) are considered special category attributes and have special legal requirements (see Section 4.1.2). Conducting this work in a lab setting allowed full control when compared to conducting an online study. Finally, the lab study facilitated an audio

recording of the session where participants were able to “think aloud” as they got results from algorithmic profiling. This method provides a detailed view of participant reactions and how they may change as they see different profiles or explanations. Conducting the experiment via a survey, for instance, would not allow for this level of granularity and authentic response. Ethics approval was obtained by submitting the appropriate paperwork - this project has approval reference number CS_C1A_19_035.

4.1.1 Recruitment

All recruited users were required to be at least 18 years of age and have at least 200 English tweets. 200 tweets was chosen as the minimum based on exiting literature (e.g. [Arnoux et al., 2017]). Flyers were sent out to various departmental mailing lists. Participants were also recruited over social media groups related to the university and city.

4.1.2 Special Category Data

Special category data required additional precautions. All volunteers were asked to consent to these profiles separately and informed that their compensation would not change if they opted not to be profiled on them. In addition to completing a checkbox on the consent form, participants had to check a box on the website to confirm that profiles would be generated on sensitive data. If participants withdrew this consent, they could hide these sensitive profiles¹ and continue with the rest of the experiment. Upon conclusion of the lab study all sensitive profiles were immediately deleted. Any tweet objects which carried special category scores were cleaned so that special category data was set to a default but all other data was retained. These steps were taken to ensure that no special category data was inadvertently made available.

In reporting on special category data, this paper uses only broad terms and generalizations to protect participant attributes. No direct quotations are reported. If a volunteer made changes to alter their profile on a sensitive attribute, a characterization of the changes (i.e. “the participant changed a substituted a suggested synonym to change a special category profile”) rather than a direct quote or explanation is used.

¹This could be achieved by navigating back to the first page and restarting without checking the consent box, see Section 4.2 for more details on what this looked like.

4.1.3 Experiment Overview

For the lab study, volunteers were asked to come to the Computer Science Department. Once there, they were asked to read an information sheet and sign a consent form.² These documents outlined the data gathered by the experiment, how it would be used, and confirmed that participants consented to an audio recording of the session. I was available throughout this process to answer any questions participants might have. After participants signed the consent form, the lab work began and lasted about forty minutes. A few general questions were asked,³ but users were mostly left free to explore the prototype as they chose. Participants were asked to “think out loud” and to share all thoughts and impressions. All participants had access to all profiles and used the same interface.

Upon completion, participants were given a £15 Amazon gift card. Participants were compensated for their time to incentivize them to volunteer for the study and to remain engaged during lab work. Once the volunteers had left, the audio recording was transcribed and the original recordings were deleted. Transcriptions both make it harder for participants to be re-identified and facilitated searching through interview content. All data except for special category data, including the gathered tweets, any edits made to them, and profiles were kept for analysis and were deleted shortly before submission of this report.

4.2 User Interface

Participants were asked to interact with the website interface on a departmental desktop in the interview room.⁴

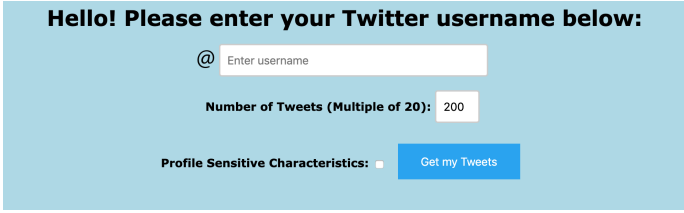
A screenshot of a web form titled "Hello! Please enter your Twitter username below:". The form is set against a light blue background. It contains a text input field with a placeholder "@ Enter username". Below this is a label "Number of Tweets (Multiple of 20):" followed by a text input field containing the number "200". At the bottom left, there is a label "Profile Sensitive Characteristics:" followed by a small square icon. To the right of this is a blue button with the text "Get my Tweets".

Figure 4.1: Mockingbird website homepage. Tweet limits are required to be a multiple of twenty due to the **Twint** library.

²These documents are included with this report.

³See questions_outline.docx included with this report.

⁴All examples shown here and elsewhere are from a test account with fake tweets. Each test tweet is a line from a poem or literary quote. These are meant to illustrate how the system works without revealing any personal information.

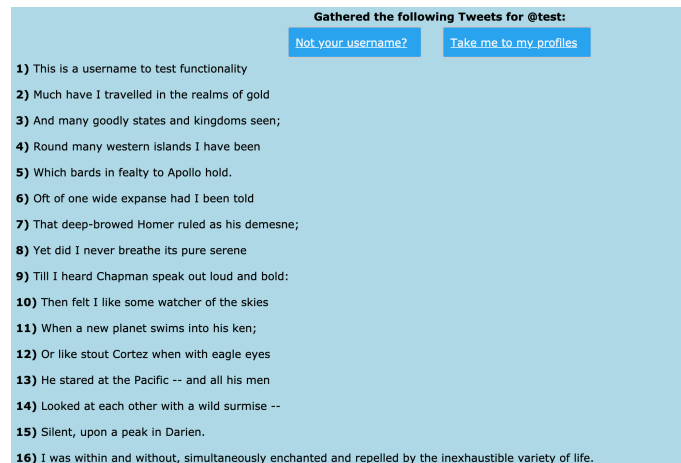


Figure 4.2: Mockingbird tweets page.

The prototype is intended to be as intuitive and simple as possible. The homepage (Figure 4.1) is a form where users submit their username, the maximum number of tweets they would like gathered, and whether or not they would like to be profiled on sensitive attributes. By default, users are opted out from sensitive profiles and the maximum number of tweets is set to 200. Upon clicking “Get my Tweets,” tweets for the provided username are scraped using *Twint*⁵ and saved as a csv. Only original tweets, including replies, are gathered (not retweets). The tweets saved in this csv are then loaded into a SQLite database managed by Django for use throughout the project. Once all tweets are loaded, the user is taken to a page displaying them to ensure that the correct account has been scraped (Figure 4.2). While tweets were being scraped, I asked users a few introductory questions regarding their expectations. This generally took under five minutes.

From here, users are taken to a page which displays all of their profiles. Profiles are grouped by attribute. Some basic information about each classifier is provided, as well as its predicted label and its confidence. For lexical models, the highest weighted words are also provided. Under each classifier is a “Find Out More” button which takes the user to the information page for each profile. At the bottom of the profiles page is a button to reset all tweets (Figure 4.5). This deactivates all current tweets and reads tweets again from the csv saved for that user. This option allows users to start over fresh after making some edits to their tweets. The Big Five personality insights are also displayed slightly differently due to how they are linked and the lack of available explanations. Some excerpts from the profile page are shown in Figures 4.3, 4.4, and 4.5.

⁵The original project is at <https://github.com/twintproject/twint>. A slight modification was required for it to run with Django; the altered code is available in the `src` directory included with this report.

We were able to generate the following profiles for you:

Sentiment

Classifier Type: Lexicon Predicted Sentiment: Positive Prediction Confidence: Low Most Negative Words: 1) sad 2) my 3) lifeless 4) regretful 5) died Most Positive Words: 1) - 2) . 3) love 4) the 5) you	Classifier Type: Pattern Predicted Sentiment: Positive Prediction Confidence: Moderate Most Negative Words: 1) sad 2) chilling 3) cold 4) down 5) worse Most Positive Words: 1) love 2) beautiful 3) many 4) loved 5) more	Classifier Type: Naive Bayes Classifier Predicted Sentiment: Positive Prediction Confidence: High Find Out More
Find Out More	Find Out More	

Gender

Figure 4.3: Mockingbird profile page.

Income

Classifier Type: Neural Network Regression Predicted Income: £22,689 Prediction Confidence: N/A Find Out More	Classifier Type: Neural Network Predicted Income (Categorical): Below Average Prediction Confidence: High Find Out More
---	---

Relationship

Classifier Type: Neural Network
Predicted Relationship: Available
Prediction Confidence: Medium
[Find Out More](#)

Politics

Classifier Type: Neural Network
Predicted Politics: Non-Political
Prediction Confidence: High
[Find Out More](#)

Figure 4.4: Mockingbird profile page.

Big Five Personality

Classifier Type: IBM Personality Insights
Predicted Openness Percentile: 99 (Open)
Predicted Conscientiousness Percentile: 34 (Not Conscientious)
Predicted Extraversion Percentile: 37 (Introverted)
Predicted Agreeableness Percentile: 9 (Not Agreeable)
Predicted Neuroticism Percentile: 83 (Neurotic)

[Find Out More](#)

[Take me to my Tweets](#)
[Reset all Tweets](#)

Figure 4.5: Mockingbird profile page.

Users were allowed to explore this page for as long as they desired and encourage to click “Find Out More” for interesting profiles. When participants click this button, they are taken to an information page for the relevant profile. This page provides a brief explanation of what the selected profile means and how it was calculated, along with a graphic showing the user’s classification. This display varies based on the classifier that is being explained (Figures 4.6, 4.7, and 4.8), but is meant to convey some notion of confidence and where the user falls relative to others.

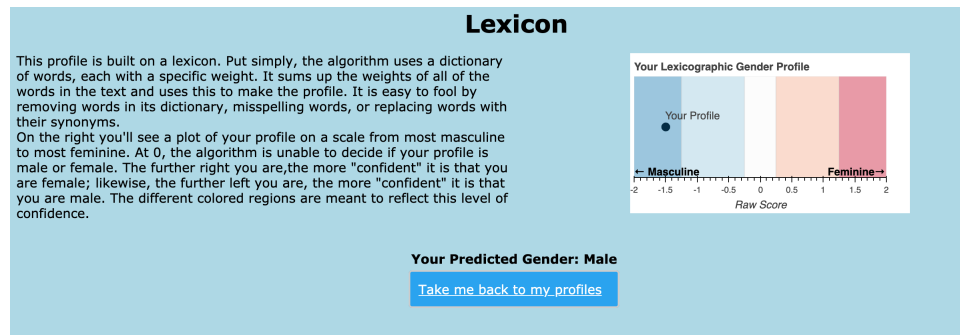


Figure 4.6: Lexicon information page (top).

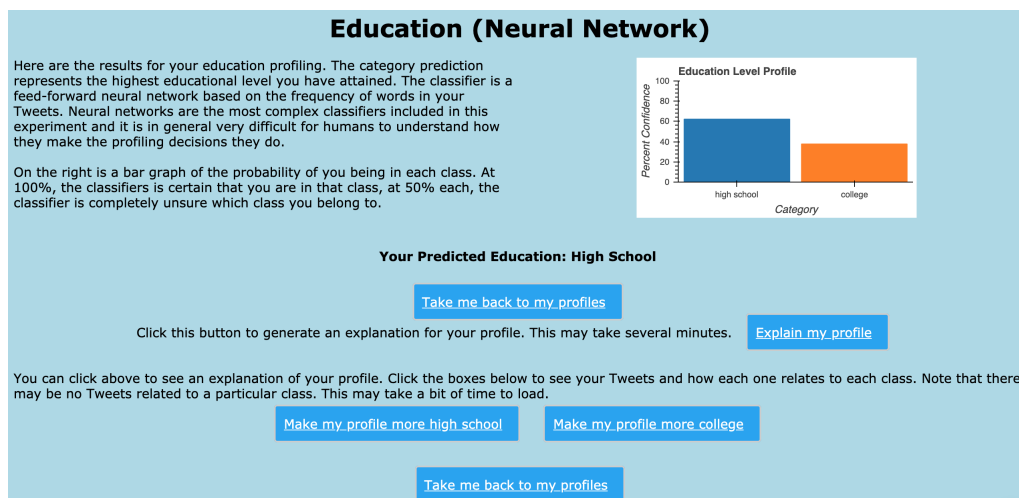


Figure 4.7: Neural network information page.

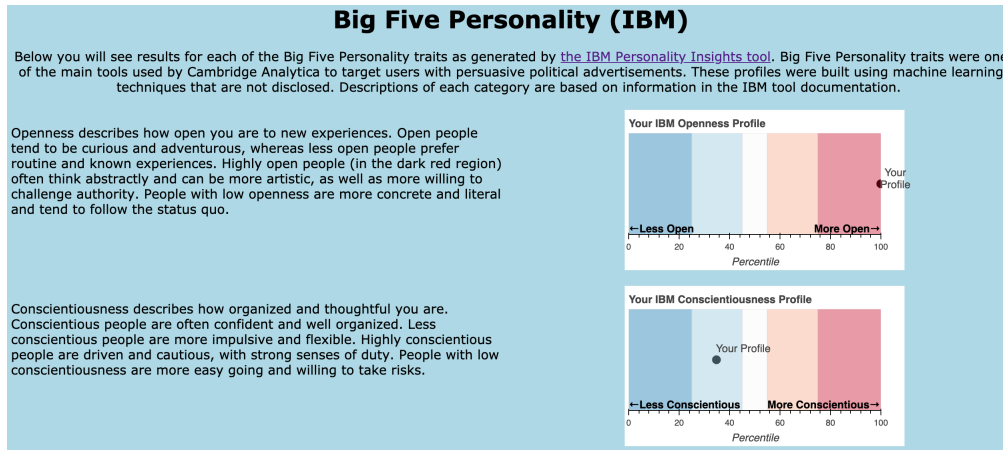


Figure 4.8: Big Five information page.

Lexical classifiers have an additional section which shows the highest contributing words along with their weights. The weights displayed are based on both the weight of the word in the lexicon and its relative frequency in the user’s tweets. This is why sometimes the highest contributing words are English stop words like “and” and “of” - these likely have small weights in the lexicon but are relatively common in the analyzed text. The decision to display explanations in this way fits with the design of the experiment: the information page for each classifier is meant to show broad, holistic information about the profile while the tweet editing page suggests small, specific changes. Also shown are the top five tweets in each category, with the two most explanatory words bolded. This information is included to give users a sense of the accuracy of the lexical analysis before they seek more information on individual tweets. As this kind of analysis is not easily calculated for other profiles, it is included only for lexicon-based profiles.

For neural network profiles, users are instead presented with an “Explain my profile” button which generates a LIME explanation for that profile. This is presented as a separate button to avoid compute time generating LIME explanations when not necessary. LIME returns the twenty-five most explanatory words and their weights. It also displays all of the tweets and highlights where the relevant words appear (see Figure 4.10). Users are not able to edit from this page directly - they must go back to the profile page and choose a target attribute. Some information pages do not offer explanations as it was either not possible to use LIME to generate one (in the case of M3 profiles) or API limitations prevented an attempt (for IBM profiles).

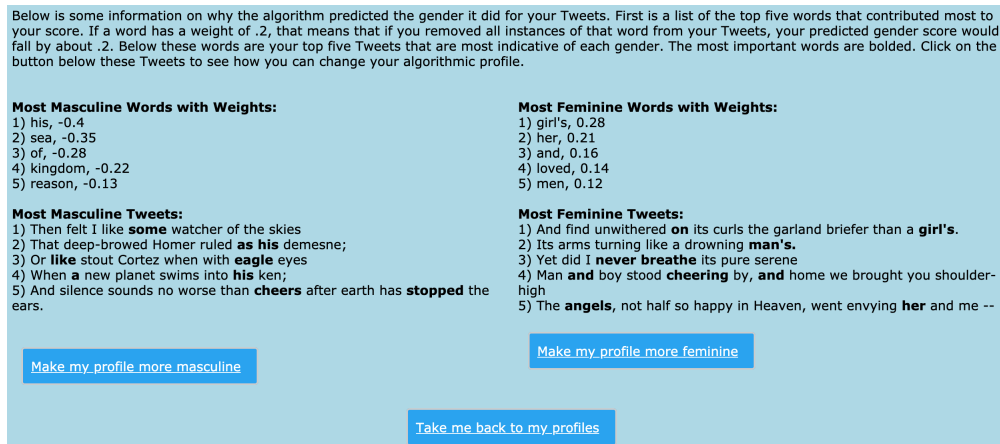


Figure 4.9: Lexicon info page (bottom).

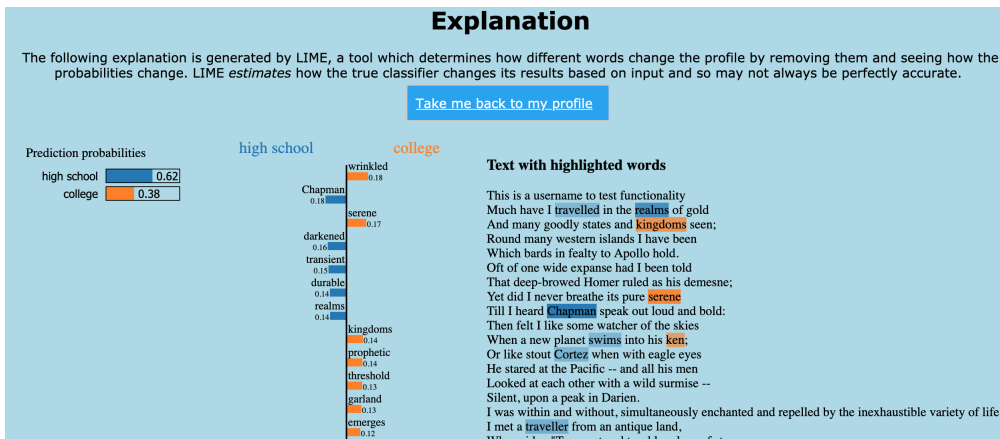


Figure 4.10: Profile explanation with LIME.

At the bottom of each information page is the option to make the profile more like one of available classes. By clicking on these buttons, users are taken to an edit page. Every edit page displays ten tweets which are least like the target class. For instance, if the target class is “young,” the edit page will show the ten tweets which are classified as the oldest by the lexical classifier. For neural networks, class confidence is used to order tweets. Each of the displayed tweets is used to auto-fill a form text box. Users are able to edit tweets directly and all changes are submitted when users navigate to a different page. Clearing out a text box effectively deletes the tweet as tweets with no text are dropped from the database. The “score” of each tweet is also shown so users can better understand how significant it is. At the top of the page is the user’s current score from that classifier. For lexicon-based classifiers, also provided are the words with the globally highest and lowest weights as well as a few tips for how to change a score.

Lexical classifiers also display up to five words which contributed most to the label and suggested synonyms where available. For the age and gender classifiers, translations are also included. For all

other classifiers, only the score and an “explain” button are available. If users click on the explain button they are taken to a new page which shows the LIME explanation of that tweet with the three most relevant words (Figure 4.13). At the bottom of the page are options to load the next ten tweets, the previous ten tweets (if not on the first page), and to return to the profiles page.

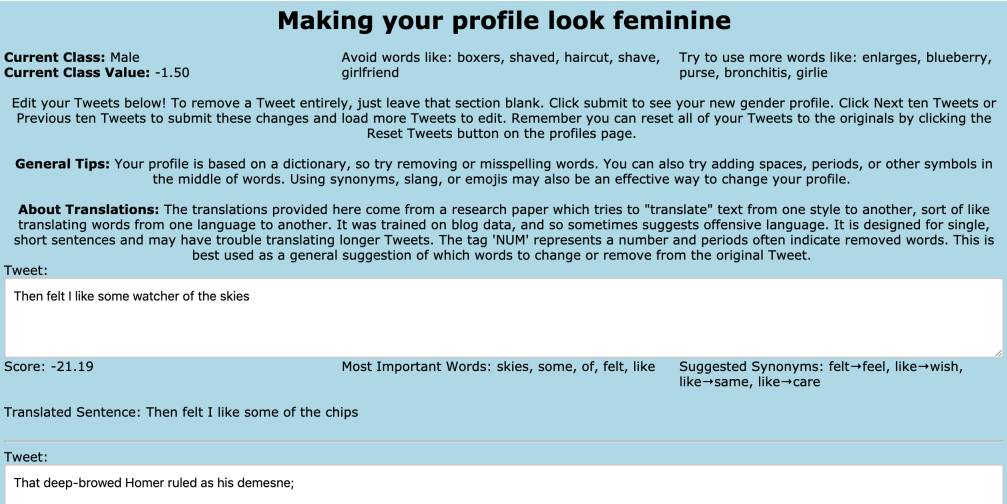


Figure 4.11: Lexicon edit page.

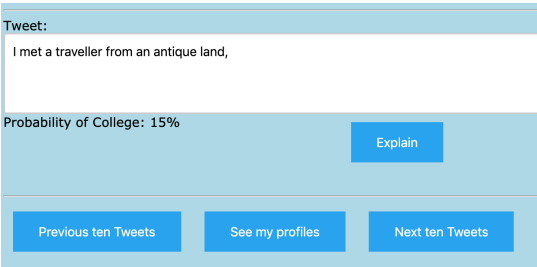


Figure 4.12: LIME edit page.



Figure 4.13: A single tweet explained with LIME.

Note that all changes to tweets are executed locally, so that the user does not actually change what is on their public Twitter account. This is logistically simpler and allows users to experiment without fear of public judgment. This does somewhat limit the generalization of these results to how people would choose to edit their public-facing data. Although users are asked to imagine making these changes to their actual profile, the stakes of social judgment are not directly present in the lab setting. An expansion of this project might allow users to “push” their final changes to their Twitter account or to check how a draft of a tweet would impact their profiles in real time.

Most users seemed to require some kind of prompting to explore changing their profiles. Frequently there would be a pause where users had viewed all profiles but not made any changes. At this point, I asked them questions regarding what they thought of profiles and encouraged them to try the editing tools. If participants experimented with editing tools naturally, I asked these questions at the end of the study. Once participants finished with editing their tweets, I asked a few questions in conclusion and handed them their gift card.

Chapter 5

Results

This chapter covers the experimental results. I first give a brief overview of the participants. Next, I outline several themes discovered in the course of the interviews and finally explain each in depth with quotations.

The themes and results below were developed from a mix of answering the questions laid out in Section 1.3 and from grounded theory - coding trends and patterns from user responses to gain new insights without necessarily using the results to support or disprove an existing theoretical framework [Charmaz and Belgrave, 2007]. By using both approaches in conjunction, I hope to satisfactorily answer the research questions as well as report unexpected results gained from speaking with participants.

5.1 Participant Information

Pseudonym	Gender	Age Range	Student	Representative
Nick	Male	20-25	Yes	No
Tom	Male	20-25	No	No
Daisy	Female	20-25	Yes	Yes
Jordan	Female	20-25	Yes	No
Jay	Male	25-35	No	Yes
George	Male	20-25	Yes	Yes

Table 5.1: Participant Information

To maintain a level of anonymity, all participants were given a pseudonym based on six of the central characters of F. Scott Fitzgerald's *The Great Gatsby*: Nick, Jay, Daisy, Jordan, Tom, and George. Participants were paired with a pseudonym that matched their gender based on a random number generator; pseudonym choice does not reflect personality, age, or any other characteristics.

See Table 5.1 for basic demographic information. The column *Representative* indicates whether or not the participant initially believed that their Twitter presence was reflective of their true self.

Three of the six participants were non-white. Of the four students, three studied a field related to computer science and the fourth studied a humanities topic. Both of the non-students worked for technology companies in non-technical roles.

In general, the interviews went smoothly and there were few technical errors. One user had difficulty scraping tweets due to privacy settings. For this user (pseudonym Daisy), only three actual tweets were gathered. After looking at profiles based on these tweets, Daisy was given access to the “test” account to experiment with altering profiles. However, as the Big Five profiles required at least 200 words Daisy did not have access to these profiles. All other participants gathered at least 200 tweets, with half of them opting for 600. One participant (pseudonym George) originally generated results based on 600 tweets and then asked to recalculate results based on over 2200 tweets. The interviews tended to run over the expected time (a half hour); they ranged from 35 to 51 minutes, with an average length of 43 minutes. This does not include the time required for users to read the information sheet and sign the consent form. All six users opted to be profiled on special category data. None asked additional questions about these profiles. The lab was loosely structured so that some participants engaged with interesting profiles fully at first, looking at explanations and considering edits, while others went through the results of all of the profiles before seeking explanations. Most users (4) required prompting to explore changing tweets. Ultimately two participants chose not to edit any tweets.

5.2 Thematic Overview

Throughout the course of studying the interview transcripts, I discovered thirteen major themes, each of which can be grouped roughly into four categories. The themes (bolded) are as follows:

Expectations **Profiling expectations were unchanged** by using the Mockingbird prototype - no participants reported that their perception of algorithmically generated profiles was significantly altered by the experiment. **Perceptions of Twitter self were unchanged** by using the Mockingbird prototype, even in cases where users were exposed to information contradicting their stated initial perceptions.

Interest in Profiles I observed **variable interest in specific profiles** across users. Although some profiles generally saw more interest than others, none were considered important by all users.

Profile confidence is important to participants, but only for a few hyper-personal profiles

(namely the Big Five personality profiles). **Profile specificity affects perceived accuracy**, in that participants were most interested in profiles with specific categories but also more likely to consider them inaccurate. Participants seemed to perceive **positive and negative classes**, even though none were presented this way. For several profiles, users expressed **uncertainty in ground truth** - this was most prevalent in the Big Five and subjectivity profiles. Lastly, participants **perceived different profile difficulties**, believing that profiling certain attributes would simply be more difficult than profiling others.

Explanations In general, **explanations had little effect** on user perceptions. Any skepticism caused by explanations seemed to be local to the profile being explained. Participants relied heavily on **mental models** that are often wrong. Participants also had difficulty adjusting mental models given clarification or explanations. **Users take the blame** for certain algorithmic shortcomings, even if their rationale is not supported by the explanations they viewed.

Willingness to Change Participants are overwhelmingly **unwilling to change** their social media data to alter algorithmic profiles that are used to target advertisements. Users who do experiment with altering their social media data all **edit differently**, although they consistently rely most on explanatory words and synonyms.

5.3 Study Results

5.3.1 Expectations

Changes in Expectations

None of the six participants admitted to changing their beliefs about profiling accuracy based on this experiment, despite having varying expectations before seeing their profiles. This may be because participants focused on the elements of the experiment that met their expectations, because their expectations were not well-defined to begin with, or because they did not believe the tools used in this experiment were comparable to the state-of-the-art. For instance, George indicated initially that his profile was very representative of his person and that profiles in general could be highly accurate. After seeing his profiles, he said that they were as accurate as expected but characterized this as “not so accurate, maybe like 60 or 65% accuracy?”¹ When asked at the end of the study about profiling accuracy, George stated “I know they can get super, super accurate... So it didn’t really change my perception.” This apparent contradiction seems best explained by a belief

¹Of the twenty-one profiles generated for this user, fifteen were deemed to be accurate (71%). Only three profiles were verifiably wrong and on the other three George expressed uncertainty regarding the ground truth.

that the methods in this project are not comparable to those used in practice, in which case the participant's expectations could not have been changed regardless of profile accuracy or inaccuracy. For participant Nick expectations seem to have been less well defined. At the beginning Nick stated "I think it will be quite hard for something to tell too much about... They'll definitely get my sense of humor, I think that's probably about enough, we'll see." Nick proceeded to characterize some profiles as "scary accurate" and noted that "it got more right than it got wrong." When asked about perceived accuracy at the end, Nick concluded "some of this data is not that accurate... I don't think it's changed my expectations but I don't think my expectations were, like, shockingly low or high in the first place." Based on these examples, it appears there was some difficulty in defining or conveying expectations or perceptions of accuracy. However, as no users indicated a change, this disproves **H1.1**.

Several users remarked that the breadth of the profiles surprised them - Big Five profiles especially seemed unexpected.

"I didn't expect for it to have fields such as sentiment and subjectivity and extravert and introvert... I didn't think that it would be interested in taking those kinds of measurements... I mean, how does it use, for example, Big Five personality?" - Tom

"I wouldn't have thought they'd try [to build all of these profiles]... Because some of it's relevant to advertising but... I don't think they need all of it. So I wonder what it's used for." - Daisy

Overall these unexpected profiles seemed to do little more than generate curiosity about how they might be used by corporations.

Twitter Self vs. Real Self

Of the six participants, three believed that their Twitter account reflected them somewhat or very accurately and three believed their Twitter account reflected them inaccurately or missed important parts of their lives. None of those who believed their profile was representative changed their mind after seeing algorithmic profiles. Tom, who believed his profile represented him "very inaccurately," seemed very impressed at what algorithms could derive from so little data:

"I think it's quite interesting and surprising... how accurate it is [given that] the data for me... isn't entirely accurate but isn't 100% off?... It's quite interesting how much information it can draw from just 200 tweets... because [personality traits] are quite specific to each individual person." - Tom

This surprise did not seem to cause Tom to believe that he was subconsciously sharing more attributes on social media than he intended. Rather, it seemed to inspire a curiosity as to how the algorithms worked and how difficult it was to develop them. Notably, this is in spite of Tom

viewing various explanations and expressing disagreement with the use of certain words. It seems that the surprise at the observed accuracy outweighed the potentially confounding nature of their explanations.

Another participant, Jay, seemed to learn something that directly contradicted his perception of his profile but did not claim that his perception was ultimately changed. Before profiling, Jay said “if you were a stranger looking at my profile you’d get a gauge of who I am as an individual... I’m not, like, ‘fake’ on Facebook... what you see on Facebook is what you’re getting when you meet me.” He confirmed that he still believed this at the end of the lab, but at first disputed the profile of being in the 32nd percentile for agreeableness. Jay claimed to be extremely agreeable overall, but admitted that the profile is “accurate on Twitter... I can definitely see me being less agreeable from my social media profile.” Given how strongly Jay insisted on both his agreeableness in person and the representativeness of his profile, this appears to be a major contradiction but it did not produce much discussion or a desire to change.

Several participants seemed to have forgotten making certain tweets but were able to eventually recall the context. Such lapses do not seem to have impacted impressions - a handful of unremembered tweets seemed unimportant. If these profiles relied on more, older data the effects of unremembered tweets could be more profound. Individuals tend to change their opinions about different topics as they age and algorithmic profiles could draw unwanted conclusions based on this older data.

5.3.2 Interest in Profiles

Variable Interest

In general, there was no consensus as to which profiles were most interesting or concerning. Each participant seemed to value different profiles. In general, organizational status, age, and gender received the least interest. Participants seemed to think gender would be especially easy to determine - several noted that their bios included their preferred pronouns or similarly indicative phrases. Although this information was available to the M3 classifiers it was not used in other profiles, leading to some confusion. Participant Jay’s lexical age profile was off by about five years, leading him to admit to being “quite surprised because in my bio it says my age.” Although his M3 profile was correct, Jay admitted that people said he looked the age predicted by the lexical profile. Jordan also noted that “I often get asked if I’m older” when faced with a lexical profile about a year older than her actual age. Participants’ mental models generally corresponded to a holistic analysis of their accounts, despite my explanations that lexical profiles were based solely on tweet content.

The income profile generated by the regression was one of the most discussed. Although most participants spent time considering this attribute, only one (Nick) was most motivated to change it. For another participant (Jordan), income was one of the least important. Other users prioritized race, political activism, personality type, and education. Several users spent time on relationship status (the least accurate profiles), although only one expressed finding this profile too invasive. In general, participants spent a lot of time considering the Big Five profiles but only one seemed to care about changing them. These results support hypotheses **H1.2a**² and **H1.2c** - users cared more about personal attributes in general but priorities differ across users.

H1.2b (regarding profile accuracy) was less conclusive. For instance, relationship status was predicted incorrectly for all users but most responded indifferently. Some seemed to find it amusing.

“I don’t care how I’m portrayed on social media in terms of my relationship... I find it quite funny to be honest.” - Jordan

On the other hand, participants whose race was profiled correctly (3) largely ignored this profile. However, of those profiled incorrectly, two were particularly interested in this attribute.³ Some participants reacted strongly to being misclassified on their education while others did not seem bothered by it. It seems that accuracy was largely secondary to which profiles the participants were inherently interested in. It is also worth noting that for all cases where at least 200 tweets were gathered, gender was predicted correctly and age was accurate to within five years. The accuracy of the age prediction did not seem to influence user interest in the profile, but it may be that if either of these two most basic attributes were substantially inaccurate they would generate more interest than attributes which are considered difficult to profile.

Importance of Confidence

Although several participants commented on the confidence of classifiers, only one expressed interest in changing it outside of the Big Five profiles. Participant George wanted to decrease his subjectivity score to be more balanced between objective and subjective, seeming to think that such a profile would be more desirable to employers. While other participants did not seem to care about class confidence in general, all participants who looked at their Big Five profile (5) expressed a desire to make smaller scale changes. Most expressed a certain percentile they expected to fall in, indicating quite a granular desire for control. This result generally disproves **H3.1** - users care only about their label except for personality, where they have a quite specific target range.

²No participants expressed concern about their sentiment and only one seemed concerned about subjectivity.

³One of the two whose race was predicted incorrectly was only able to gather three tweets and so it is difficult to gain much insight from those results.

Specificity

Interest in and perception of profiles seemed to depend somewhat on the specificity offered by the profile. For instance, several participants noted that the age predicted by the lexicon was incorrect because it was off by several years but that the age predicted by the M3 classifier was correct. However, the M3 classifier breaks down age into groups of at least ten years. If instead of estimating an exact age the lexical results were converted to the same categories as the M3 classifier, in all but one case the two different techniques would agree. When the lexical age profile was exactly correct, the participant expressed considerable surprise, remarking “that’s scary actually, I’m not sure how...” (Nick). Similarly, users were quite interested in the income regression but few even looked at the categorical income results. Participants often expressed interest paired with doubt that such accurate estimates were possible:

“I like the specificity of how specific it is. It thinks I make exactly £23,378. Why does it think that? How do you know?” - Nick

“I don’t see how an AI algorithm could predict that somebody makes like 70 grand versus like 30 grand or 100. But it could say like ‘average’ or ‘above average’ or ‘below average.’” - George

It seems that in general, more specific profiles generated more interest but were received with more skepticism. Participants were curious about how the profiles arrived at the specific prediction but doubted that it could be accurate. On the occasions when these profiles were highly accurate, participants were very impressed but mostly they seemed more satisfied with broader category predictions.

Positive & Negative Classes

A recurring theme throughout the labs was that most participants implicitly considered some classes as “negative” and others as “positive.” Education was a clear example of this - participants saw being more educated as a desirable trait.

“[Twitter] would like to see the smarts. Maybe this is why I’m seeing all of the mobile game ads. Because it thinks I’m a schmuck who went to high school and likes playing mobile games. Uhh, yeah, all of the other things I’m not too bothered about.” - Nick

“[Education] is always tough because I went to college but didn’t go to uni. But my qualification, what I’m currently doing is on a degree level. But I’m kind of studying, like part time.” - Jay

Nick especially seemed quite bothered by this profile, connecting it to both income and perceived intelligence. Jay attempted to explain how, although this profile was technically accurate, he was better educated than the classifier showed. These two users were also concerned about their income

profile. Nick wanted to increase the estimate and Jay seemed uncomfortable with the profile in general. Jay looked at the information page for the income profile and, despite prompting to view the explanation, chose not to. Jay later remarked that the income profile was “pretty much close to be honest, which is pretty impressive.” It seemed that Jay was uncomfortable that the algorithm accurately profiled their income as below average. A different participant also observed that this might be demoralizing.

“Having it there all the time, saying ‘hey you sound like you have low income because of these words,’ I think that would be maybe kind of sad, having that information.” - Daisy

Other participants had value judgments about being profiled as political, subjective, and on several Big Five attributes. These occasionally conflicted - one participant saw being politically active as a positive trait whereas others saw it as negative. Additionally, some profiles that were considered to have positive or negative value by some users were perceived as value neutral to others. There did not seem to be much consensus, but users reacted more strongly to being profiled on attributes they associated with a value:

“I’m flattered that I’m considered to be agreeable. Because I like that characterization I approve of it.” - Jordan

“I’m at a high percentage [for openness] because I consider myself a very open person. I don’t like that stereotype where as guys we’re closed off and stuff. So I like breaking that stereotype.” - Jay

“Maybe I need to leave smarter tweets I guess?” - Nick

“That is actually surprising me, because I see myself as an open person. It’s got me down as less open. That’s a bit of a hit to my ego.” - Tom

“[I would like to change] the emotional part. Because that’s probably not a good thing, professionally, to be known as.” - George

Even though this varied significantly from user to user, it appeared to be one of the most important factors for user interest.⁴

Uncertain Ground Truth

For the Big Five and subjectivity profiles, many users were unsure of what their profile should be. All users profiled on Big Five admitted to being unfamiliar with it and were unable to compare the profiles to a test they had taken previously. Interestingly three of the five participants indicated that they thought they should be exactly in the middle between extraverted and introverted. Jay especially seemed to see introversion as a somewhat negative trait. Without having participants also

⁴To my knowledge, nothing about the presentation of the profiles indicated that one class was more desirable than another; this is supported by the fact that perceptions varied significantly between users, suggesting that the causes were exogenous.

take a standard Big Five survey, it is difficult to draw conclusions about this confusion. Participant desires to change their Big Five profile may be rooted in a poor understanding of the meaning of each category or reflect where participants *want* to be rather than where they are.

Perceived Profile Difficulty

As discussed earlier, users seemed to expect that some profiles would be more difficult to generate than others. Age, gender, and sentiment all seemed to be perceived as straightforward, whereas relationship status, income, and race were all seen as difficult to profile. This belief seemed to be a property of the attribute being profiled, not the complexity of the model generating the profile.

“I couldn’t tell from my tweets if I were religious or not, I don’t think a human could. And relationship, I don’t know how it does that anyway, I don’t know if you’re tagging a significant other on Twitter, then...” - Nick

“But I guess it didn’t really predict any of my income because I don’t post about my income but I don’t see how it could predict this. I don’t think you could predict this based on my tweets.” - George

George was quite hostile towards the income regression specifically.

“I’m sure they can’t get it [income] accurately for anyone else... I mean it’s just what they predict about you, you know, not what it actually is. And it isn’t like they can use it against me or have proof.” - George

Despite having a high degree of confidence in algorithmic profiles, George was convinced that income could not be accurately predicted from tweets.⁵ The reasoning for this was not clear, but it seemed as if George believed that no reasonable degree of accuracy was obtainable from tweets unless he explicitly mentioned income. In general, participants seemed to understand when these “difficult” profiles were incorrect, often attributing it to the topics they tweet about or the inherent difficulty of the task.

While the ease of profiles largely consistent across participants, some expected certain attributes to be easier to guess based on what they shared online. Even though Nick observed that he thought several profiles would be difficult if not impossible to generate accurately, he also remarked that “education is the one thing I think ‘ok, maybe it should’ve [gotten that right].’ I do talk about living in Oxford and stuff like that. I thought maybe it would pick up on that.” Nick seemed quite bothered that this attribute that was obvious to humans was missed by the algorithm. Perhaps if other seemingly “easy” profiles were also inaccurate users would respond more strongly.

⁵At the end of the study, George said that he “know[s] that they [companies like Google] can get super, super accurate.” It is unclear if he believes such companies could achieve an accurate income regression.

5.3.3 Explanations

Explanations Have Little Effect

As no expectations regarding profiling accuracy were changed, it does not appear that explanations had an explicit impact on user trust. However, all participants saw both explanatory words that agreed with their expectations and explanatory words which did not. In some cases, participants seemed to accept explanatory words even if they did not make sense, indicating a strong trust in the profiling methods. In other cases, explanations seemed to erode user trust:

“I wouldn’t think ‘hello’ or ‘woohoo’ would have a strong weighting [regarding race]... I think this is really random for predicting race.” - Daisy

“I don’t really understand why it thought I was taken because of those words.” - George

“This is how people in relationships talk... Reddit [laughs], apparently, that’s a lie.” - Nick

Participants seemed similarly disillusioned by profiles missing words they thought were obvious indicators. Jordan, after being profiled as having a high school education, noted “recently I’ve mentioned Master’s degree or DPhil or something... So, that’s not accurate.” Missing words that would seem obvious to human readers made participants feel that the model was somewhat random.

It does seem that unconvincing explanations only affected trust in the model they were explaining rather than all models of that type or all algorithmic profiles in general. Participants who complained about one profile’s explanations were just as willing to praise the next models explanations if they matched expectations. Taken together, these results do not substantially support **H2.1** - explanations seemed to have a small impact on user trust in most cases.

Mental Models

Each participant appeared to have some intuition for how algorithmic profiles are generated and some asked for explanations of how the models worked. Often these mental models were inaccurate - three users pointed to information in their bio that indicated their age or gender to explain lexical profiles which did not have access to this information. Participants adjusted their expectations frequently based on explanations:

“Why is my most masculine word wife?... I guess the bias is people tend to talk about the other gender... ‘server’ and ‘adapter,’ so that’s techy words isn’t it [that relate to high income]?.. I can guess why this one is probably low income, it’s swearing... swearing is a low-income thing.” - Nick

“I have an opinion column for a newspaper... and I talk a lot about mental health issues... So a lot of those articles I write for it are very personal in nature and relate to my life experience. And then I will often tweet them with a quote from it, which is perhaps why I have [a high neuroticism score].” - Jordan

“If you tweet a football result that is objective but if you tweet about performance, that’s subjective isn’t it?” - Tom

Some of these adjustments, such as those mentioned by Nick, are informed by explanations. Others are based purely on profiles - even when explanations were available users seemed to make their own assumptions and inferences. Some degree of this was necessary as the explanations do not explain *why* certain words are relevant. It is left to the users to make this determination and judge its sensibility. The reliance on theories even in the face of explanations indicates that they may be quite pervasive. Short of fully understanding how the model works, humans fill in the gaps with logic which may or may not be true. This realization supports the case for fully interpretable models: if a model is interpretable at every step, there are no gaps to be filled with human reasoning. All intuitions can be immediately verified and there is no risk of misunderstanding.

Participant Explanations of Profiling Inaccuracy

Participants frequently offered explanations for algorithmic shortcomings based on their use of social media rather than a failure of the algorithm.

“High school, eh, I’m doing a Master’s degree!... I have had the Twitter profile for a while, I tweeted a lot when I was younger and I tweet a lot now, and there was a big period in the middle where I didn’t use it at all. So maybe that’s why?” - Nick

“It’s got my age wrong, but also as well a lot of my tweets are from when I was younger.” - Tom

“I wonder if [my inaccurate relationship profile] is because I interact a lot with my ex-girlfriend on Twitter. Or I’ll say I love someone quite a bit because I’m quite open about loving my friends.” - Jordan

“[In response to an incorrect relationship profile] Yeah, I don’t talk about my relationships that much.” - Jay

While these explanations may be generally correct they seem to “forgive” algorithmic errors rather than decrease user trust. The algorithm may have predicted a younger age based on old tweets but it would be logical for the algorithm to anticipate this and give more weight to recent tweets. Rather than question the wisdom of the algorithm, participants seemed to attribute its failures to their own actions. This is consistent with results found in [Eslami et al., 2018] where participants were observed to “take the blame” for algorithmic shortcomings.

5.3.4 Willingness to Change

Unwillingness to Change

Five participants claimed at the end of the study to ultimately be indifferent about algorithmic profiles and the sixth seemed unconcerned with the current state of algorithmic profiling.

“I don’t really care about the amount of data they collect or what they know about me. To be honest.” - George

“I don’t really care that much because it’s just adverts.” - Daisy

“On a personal opinion, I’d say it doesn’t bother me that much to be honest.” - Jay

“I don’t care all that much about how I’m portrayed on social media, to the worry of my family members.” - Jordan

“In my opinion, as long as my information isn’t being given out to people who might use it to commit fraud... then it doesn’t bother me that much.” - Tom

“It’s made me feel a bit more confident that some of this data is not that accurate.” - Nick

Due to this apathy, participants saw profiles as interesting but were unwilling to make changes other than out of curiosity. This unwillingness to change appears to be strongly held.

“I wouldn’t be super willing to change it a ton because then it kind of isn’t me any more?... If you want to say something you should say it.” - Daisy

“I feel like people shouldn’t repress themselves when they post about something. I feel like if you want to say something, just say it, right?” - Jay

“I would just carry on how it is... I use social media for the way I want to use social media sort of thing.” - Tom

Participants seemed to view editing their social media data as either dishonest or running counter to their reasons for using the platform. Two users stated that they would be more likely to make a new account or quit using the platform than edit tweets if they were concerned about algorithmic profiles. Only one participant, George, was really open to change, claiming that he’d “probably be willing to change anything” to control algorithmic profiles although he also expressed apathy regarding data collected for advertising. George eventually decided that he would be likely to delete “harmful” tweets in certain situations rather than try to edit their content. Overall, participants in this study were not interested in changing the contents of their tweets to confuse algorithmic profiles, largely invalidating **H3.2**.

Methods of Editing

Although users expressed an unwillingness to change their actual tweets, several experimented with editing tools locally. There appeared to be no commonality between how users decided to edit tweets. One participant simply replaced explanatory words with the suggested synonyms, ignoring sense and grammatical correctness.⁶ Another participant deleted entire tweets and attempted to reword them from scratch, omitting the explanatory words. A third participant deleted some explanatory words and replaced others with synonyms before resorting to simply deleting about a dozen tweets. A fourth deleted vowels or caused misspellings of the explanatory words.

⁶This user was editing tweets from the test account as too few of their own could be gathered. It is reasonable to assume they may have behaved differently with their own data.

No participant appeared to use the “translations,” although two remarked on the errors they made. They appeared to be of too low quality to be useful. No users made use of the “global” maximum and minimum words either. All focus appeared to be on synonyms or removing words altogether. This suggests that having accurate explanations is more useful than having complex tools to help edit tweets.

Chapter 6

Conclusion

In this section, I first outline the major limitations of this work. I then conclude the results of this study and discuss several implications for future work. Finally I provide some closing remarks.

6.1 Limitations

The major limitation of this project is the small, unrepresentative sample used for the study. To gather more accurate results, many more users of different demographics and backgrounds would need to be recruited. Rather than conducting a lab study, a survey could be set up online to gather a sufficient number of users responses. Additionally, the data used to train many of the classifiers was noisy and potentially unreliable. Building models on better data that is more representative of Twitter and with more accurate labels could greatly boost performance. Adding support for languages other than English would be an important step; this might be accomplished by using automated translation software as suggested in [Wang et al., 2019]. By including additional social media networks such as Facebook, LinkedIn, and Instagram, more holistic profiles could be generated and the tool opened up to more social media users.

Another limitation was the attributes and techniques used. With different datasets that include entire tweets rather than just bag-of-words data, more complex neural networks could be employed. This could greatly improve results and is likely to be more reflective of the industry. Explanations and editing tools focused on outdated methods of profile generation could potentially still be useful but in the worst case they could be counter-productive. Although one would expect highly-weighted words in a lexicon to be similarly important to a deep learning model there is no guarantee that this would be the case. More complex models are likely to discover more complex semantics (including negation and sarcasm) which are lost to simpler models. Even though the models used in practice

cannot be known for certain, keeping tools up-to-date both with data that are reflective of the current milieu of social media and the state-of-the-art profiling techniques ensures that they will continue to be useful.

6.2 Conclusions

In this project, I have first observed the particular risk posed by algorithmically generated profiles and the lack of understanding of these profiles and risks by most social media users. To address this problem, I created a novel prototype tool called Mockingbird which uses Twitter data to profile individuals on over twenty attributes, provides users with explanations of these profiles, and allows them to edit their social media data to alter their profiles. Mockingbird also provides a pair of editing tools to assist users in deciding how to alter their social media data. To test the Mockingbird prototype, I conducted a lab study which allowed users to experiment with this tool. I now return to the main research questions outlined in Section 1.3.

6.2.1 Addressing Research Questions

Based on the results of these lab sessions, it does not appear that algorithmic profiles are important to social media users (**RQ1**). Even profiles which participants found interesting or surprising did not cause a sense of discomfort or a desire to change them. Participants were generally curious, but did not seem to strongly relate these profiles to either how they were perceived by other Twitter users or potentially problematic usage of them by a third party. A degree of this indifference can be attributed to both an ignorance of potential uses of algorithmic profiles and an apathy regarding the usage of profiles for targeted advertisements.

Similarly, participants seemed largely unaffected by explanations (**RQ2**). Explanations that did not match user expectations did not decrease overall confidence or trust in algorithmic profiles. Occasionally, explanations were ignored. One user, upon not seeing the words they expected as explanations for their political profile, searched for these words in their tweets anyway and pointed them out to explain their label. Most participants saw explanatory words they disagreed with but this did not seem to undermine their overall trust in the model. Perhaps user trust would have been more greatly impacted if participants were made to view *all* profile explanations. As it was, users saw only explanations for profiles they were interested in. In many cases, these could be profiles participants thought would be difficult to generate or believed were incorrect. In these cases, seeing explanations providing insight into inaccurate models may not have been surprising. It may have been more effective to see potentially illogical explanations for profiles users thought were accurate.

Future work could experiment with different types of explanations, as in [Binns et al., 2018].

In general, there seemed to be little interest in altering tweets to change profiles (**RQ3**). This is likely due to the perceived unimportance of ad profiles. Participants had trouble seeing why their profiles, even on sensitive characteristics, were problematic if they were used only for ads. One user noted that they blocked advertisements so they would not even be able to see the impact of these profiles, while another was in favor of accurate ad profiles so long as their information was not being used for fraudulent purposes.¹ As controlling algorithmic profiles requires effort and may come at some perceived cost (see Section 5.3.4), most users were simply unwilling to do it for the purposes of fooling advertisers. However, that does not mean that such techniques could not be applied in higher-stakes situations. Three participants expressed concern about employers viewing their Twitter account; one of these kept their profile on private to prevent this. While discussing other potential usages of algorithmic profiles, one user brought up a US immigration policy.

“The Trump administration announced recently that they will ask for your Twitter profiles to decide whether they will give you a visa or not²... I would definitely delete stuff based on tweets I think... could have any consequences... [this tool] would be great, yeah.” - George

While other users mentioned creating a new account or moving away from Twitter rather than adjusting their usage of it, this statement does indicate interest in altering current profiles in certain higher stakes applications. Users simply do not seem to care about their algorithmic profiles when they are being used only for advertising purposes. An expansion of this project could focus on how a handful of algorithmically generated profiles contribute to a simulated high-stakes situation. For instance, a profile about political activism could be used to “determine” visa status. Participants would then only deal with this profile directly and would be faced with a narrower range of available edits to alter this algorithmic decision. By recontextualizing algorithmic profiles in this explicitly high stakes situation, participants would have a stronger motivation to change their profile and face potentially difficult trade-offs between altering their data to influence profiles and remaining true to their original posts.

Another expansion of this project could display to users a number of potential consequences of being profiled a particular way in terms of high-stakes applications. For instance, it could claim that having a below average income profile made an individual 50% less likely to be approved for a loan or a political profile made them ineligible for a visa. If done correctly, this expansion would force users to consider the potential harm done by their algorithmic profiles and encourage them to

¹This second participant, Tom, contended that the only profiles he would consider invasive would be ones based “the type of porn I watch.”

²See <https://www.nytimes.com/2018/03/30/world/americas/travelers-visa-social-media.html> [Chan, 2018]

make changes.

Recruiting users with at least 200 tweets may also have influenced this unwillingness to change. People who tweet frequently may feel more comfortable with their persona online and the way they are perceived. This might make them less likely to trust algorithmic profiles or alter their tweets. Users who “lurk” on Twitter, that is, have an account but post less frequently, might be more open to making changes. These users may be less confident in what their Twitter profile says about them or more nervous about how they are perceived online. This could make them more willing to trust algorithmic profiles or to see the importance of controlling them. This is speculation and would require further research.

In conclusion, giving users the ability to see their algorithmic profiles did not change their perception of algorithmic profiles or prompt them to alter their behavior on social media. Recent developments suggest that individuals who opt to quit using the platform or make a “professional” public account may in fact be better off than those who try to edit existing profiles. On 27 August 2019, a freshman travelling to the United States to begin their undergraduate education was denied entry due to the content of his friend’s posts [Zraick and Zaveri, 2019].³ If social media users are made responsible not only for their own content, but also for the content of those that they interact with online, perhaps moving away from platforms under scrutiny or separating public-facing accounts from private ones is the best course of action. The impact of social media usage on individual lives is constantly growing and developing. It is vital that users are informed of these potential consequences and provided with tools to help manage them.

6.2.2 Implications for Future Work

This work showed that users tended to focus only on explanatory words and their synonyms. Improving existing tools or adding new ones could increase the ability of users to alter their profiles. A tool which automatically suggests minor changes to confuse tokenizers as in [Ebrahimi et al., 2017] and [Hosseini et al., 2017] could be a useful addition. Instead of focusing on one attribute, several could be manipulated at once, potentially preventing situations in which correcting for one target attribute unintentionally changes another. Further work on style “translation” as in [Shetty et al., 2018] is particularly interesting. Perhaps the quality of translations could be improved by training directly on recent Twitter data or by crowdsourcing paired stylistic translations. A⁴NT networks could be trained on more attributes or even on groups of attributes. A tool that translates from a given tweet to a more narrowly specified class (such as “young liberal man”) would allow users more control over the style of translation. This would come at the cost of training many more models and gathering

³<https://www.nytimes.com/2019/08/27/us/harvard-student-ismail-ajjawi.html>

more data. Even with more targeted translations, user approval would still be necessary to avoid undesirable changes. Synonym suggestions could be improved by using a more complex tool than WordNet and by incorporating synonym suggestion with LIME explanations. The process could work as follows: first, LIME generates a number of explanatory words. Then, candidate synonyms are generated for these words. Next, using a metric such as language loss (see Equation (2.12)) synonyms which do not make semantic sense are eliminated. Finally, sentences with candidate synonyms are classified and those closest to the target class are suggested to the user. This process would generalize synonym suggestion beyond lexical classifiers as well as reduce the appearances of nonsensical or confusing suggestions.

Another avenue of future research would be to turn Mockingbird into a real-time tool which showed the effect of a potential post. The prototype could be turned into a browser extension which highlights posts in real time, making it easy for users to actually edit their social media data. The extension could also analyze drafts of social media posts to see how they are likely to impact various profiles and suggest changes *before* posting. In Goffman's terms, this would allow "actors" to effectively simulate a performance and receive feedback on it before they commit to a final version. Such interactive tools require a smaller time commitment than combing through years of posts and would allow users to hone in on particularly salient attributes for their personal identity. All participants expressed interest in this, although most doubted it would change the content of their tweets. George was the exception to this, indicating that he would be willing to change his tweets using such a tool to alter his political or relationship profile.

A final expansion of this project would be to develop an iterative approach to profiling and user changes. As social media companies gain insight into what people are or are not willing to change to alter algorithmic profiles, they may attempt to alter their algorithms accordingly. These alterations may depend on the application relevant to the profiling company and how accurate they believe their profiles are. One company might look to prioritize features users are willing to change as they think user changes will make profiles more accurate and improve the company's bottom line. Another company may think that user changes tend to deceive or confuse their algorithms and so increase focus on elements users are less willing to change. A future project could design profiling tools aware of user preferences for changing their tweets. It is likely that users will be frustrated by profiles they find too difficult to change but might believe that more malleable profiling tools are too sensitive. Such findings may be useful to social media companies as they adapt to increased user awareness of and control over algorithmic profiles.

6.3 Closing Remarks

Overall, this project was successful in creating the Mockingbird prototype, providing a valuable tool to view, understand, and manage algorithmically generated profiles. It also succeeded in addressing the stated research questions, even though users did not display a willingness to incorporate insights or tools from Mockingbird into their social media behavior.

The results of this project highlight the need for greater user awareness of algorithmically generated profiles and their potentially life-altering impacts. I believe that participant response would have been substantially different if they were more conscious of the serious ramifications their profiles could have. The one participant who was aware of this, George, was also the one most willing to change their social media to control algorithmic perceptions. Without greater awareness, algorithmically generated profiles will continue to be developed and deployed. It seems likely that governmental regulation will be required to protect users by mandating systems which adhere to fairness, accountability, and transparency standards.

This project suggests that user education is at least as important as providing users with tools to control algorithmic profiles. Without a reason to use them, tools gather dust; unless social media users can be convinced that their algorithmically generated profiles can affect their lives in meaningful ways, any attempts at empowering them to alter these profiles will be met with a reluctance to enact change.

References

- [Ali et al., 2019] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. (2019). Discrimination through optimization: How facebook’s ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095*.
- [Andreou et al., 2018] Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of facebook’s explanations. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.
- [Angelino et al., 2017] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44. ACM.
- [Angwin and Parris Jr, 2019] Angwin, J. and Parris Jr, T. (2019). Facebook lets advertisers exclude users by race.
- [Angwin et al., 2019] Angwin, J., Varner, M., and Tobin, A. (2019). Facebook enabled advertisers to reach ‘jew haters’.
- [Arnoux et al., 2017] Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., and Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. In *Eleventh International AAAI Conference on Web and Social Media*.
- [Azucar et al., 2018] Azucar, D., Marengo, D., and Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159.
- [Binns et al., 2018] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). ‘it’s reducing a human being to a percentage’: Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 377. ACM.
- [Brennan et al., 2009] Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40.
- [Cadwalladr and Graham-Harrison, 2018] Cadwalladr, C. and Graham-Harrison, E. (2018). The cambridge analytica files. *The Guardian*, 21:6–7.
- [Chan, 2018] Chan, S. (2018). 14 million visitors to u.s. face social media screening. *The New York Times*.
- [Charmaz and Belgrave, 2007] Charmaz, K. and Belgrave, L. L. (2007). Grounded theory. *The Blackwell encyclopedia of sociology*.
- [Chen et al., 2014] Chen, J., Hsieh, G., Mahmud, J. U., and Nichols, J. (2014). Understanding individuals’ personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414. ACM.

- [Cheney-Lippold, 2011] Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, 28(6):164–181.
- [Chollet et al., 2015] Chollet, F. et al. (2015). Keras. <https://keras.io>.
- [Coltheart, 1981] Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.
- [Confessore and Hakim, 2017] Confessore, N. and Hakim, D. (2017). Data firm says ‘secret sauce’ aided trump; many scoff. *The New York Times*, 6.
- [Conover et al., 2011] Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE.
- [Culotta et al., 2015] Culotta, A., Kumar, N. R., and Cutler, J. (2015). Predicting the demographics of twitter users from website traffic data. In *AAAI*, pages 72–78.
- [Dewey, 2016] Dewey, C. (2016). 98 personal data points that facebook uses to target ads to you.
- [Dickerson et al., 2014] Dickerson, J. P., Kagan, V., and Subrahmanian, V. (2014). Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 620–627. IEEE Press.
- [Dos Santos and Gatti, 2014] Dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.
- [Duhigg, 2012] Duhigg, C. (2012). How companies learn your secrets.
- [Ebrahimi et al., 2017] Ebrahimi, J., Rao, A., Lowd, D., and Dou, D. (2017). Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.
- [Ehsan et al., 2019] Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. (2019). Automated rationale generation: a technique for explainable ai and its effects on human perceptions. *arXiv preprint arXiv:1901.03729*.
- [Emmery et al., 2018] Emmery, C., Manjavacas, E., and Chrupała, G. (2018). Style obfuscation by invariance. *arXiv preprint arXiv:1805.07143*.
- [Eslami et al., 2016] Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., and Kirlik, A. (2016). First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2371–2382. ACM.
- [Eslami et al., 2018] Eslami, M., Krishna Kumaran, S. R., Sandvig, C., and Karahalios, K. (2018). Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 432. ACM.
- [European Union, Parliament and Council, 2016] European Union, Parliament and Council (2016). General data protection regulation. *Official Journal of the European Union*, L 119/1.
- [Farnadi et al., 2016] Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F., and De Cock, M. (2016). Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3):109–142.
- [Garcia, 2016] Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4):111–117.

- [Gilpin et al., 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE.
- [Goffman et al., 1978] Goffman, E. et al. (1978). *The presentation of self in everyday life*. Harmondsworth London.
- [Golbeck and Hansen, 2011] Golbeck, J. and Hansen, D. (2011). Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1108. ACM.
- [Golbeck et al., 2011] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.
- [Goodfellow et al., 2014a] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- [Goodfellow et al., 2014b] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- [Gou et al., 2014] Gou, L., Zhou, M. X., and Yang, H. (2014). Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 955–964. ACM.
- [Hendricks et al., 2018] Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models (extended abstract). *CoRR*, abs/1807.00517.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- [Hogan, 2010] Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386.
- [Hosseini et al., 2017] Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.
- [Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- [Jang et al., 2016] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.
- [Jia and Liang, 2017] Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.
- [Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [Kiritchenko et al., 2014] Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

- [Kosinski et al., 2013a] Kosinski, M., Stillwell, D., and Graepel, T. (2013a). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.
- [Kosinski et al., 2013b] Kosinski, M., Stillwell, D., and Graepel, T. (2013b). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.
- [Kramer et al., 2014] Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- [Lepri et al., 2018] Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627.
- [Lin et al., 2017] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- [Mairesse et al., 2007] Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.
- [Matz et al., 2017] Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719.
- [McCrae and John, 1992] McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.
- [Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- [Mittelstadt et al., 2016] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.
- [Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.
- [Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- [Ozer and Benet-Martinez, 2006] Ozer, D. J. and Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57:401–421.
- [Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.
- [Park and Grow, 2008] Park, J. S. and Grow, J. M. (2008). The social reality of depression: Dtc advertising of antidepressants and perceptions of the prevalence and lifetime risk of depression. *Journal of Business Ethics*, 79(4):379–393.
- [Pennacchiotti and Popescu, 2011] Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- [Pennebaker et al., 2015] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

- [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- [Poursabzi-Sangdeh et al., 2018] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.
- [Preoŭciuc-Pietro et al., 2015a] Preoŭciuc-Pietro, D., Lamos, V., and Aletras, N. (2015a). An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764.
- [Preoŭciuc-Pietro et al., 2015b] Preoŭciuc-Pietro, D., Volkova, S., Lamos, V., Bachrach, Y., and Aletras, N. (2015b). Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.
- [Quercia et al., 2011] Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE.
- [Rao et al., 2015] Rao, A., Schaub, F., and Sadeh, N. (2015). What do they know about me? contents and concerns of online behavioral profiles. *arXiv preprint arXiv:1506.01675*.
- [Rasmussen, 2003] Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.
- [Reddy and Knight, 2016] Reddy, S. and Knight, K. (2016). Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.
- [Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.
- [Rocher et al., 2019] Rocher, L., Hendrickx, J. M., and de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1).
- [Rudin, 2018] Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*.
- [Sap et al., 2014] Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.
- [Shetty et al., 2018] Shetty, R., Schiele, B., and Fritz, M. (2018). A4nt: author attribute anonymity by adversarial training of neural machine translation. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1633–1650.
- [Smedt and Daelemans, 2012] Smedt, T. D. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.
- [Smith and Kidder, 2010] Smith, W. P. and Kidder, D. L. (2010). You’ve been tagged!(then again, maybe not): Employers and facebook. *Business Horizons*, 53(5):491–499.
- [Stumpf et al., 2016] Stumpf, S., Bussone, A., and O’Sullivan, D. (2016). Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

- [Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- [Sweeney, 2013] Sweeney, L. (2013). Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*.
- [Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- [Ur et al., 2012] Ur, B., Leon, P. G., Cranor, L. F., Shay, R., and Wang, Y. (2012). Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security*, page 4. ACM.
- [Ustun et al., 2019] Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM.
- [Veale et al., 2018] Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, page 440. ACM.
- [Volkova et al., 2015] Volkova, S., Bachrach, Y., Armstrong, M., and Sharma, V. (2015). Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- [Wang et al., 2019] Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Jurgens, D., et al. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067. ACM.
- [Warshaw et al., 2015] Warshaw, J., Matthews, T., Whittaker, S., Kau, C., Bengualid, M., and Smith, B. A. (2015). Can an algorithm know the real you?: Understanding people’s reactions to hyper-personal analytics systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 797–806. ACM.
- [Yang and Li, 2013] Yang, H. and Li, Y. (2013). Identifying user needs from social media. *IBM Research Division, San Jose*, page 11.
- [Youyou et al., 2015] Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.
- [Zhao et al., 2017] Zhao, Z., Dua, D., and Singh, S. (2017). Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.
- [Zheng et al., 2018] Zheng, X., Han, J., and Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.
- [Zhu et al., 2014] Zhu, X., Kiritchenko, S., and Mohammad, S. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 443–447.
- [Zliobaite, 2015] Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*.
- [Zraick and Zaveri, 2019] Zraick, K. and Zaveri, M. (2019). Harvard student says he was barred from u.s. over his friends’ social media posts. *The New York Times*.