YaleNUSCollege

GHOSTS IN THE MACHINE: APPLYING & EVALUATING EXPLAINABLE AI TECHNIQUES IN LEGAL DECISION MAKING

TRISTAN KOH LY WEY

Capstone Final Report for Yale-NUS BSc (Honours) in Mathematical, Computational and Statistical Sciences & LLB (Honours) Double Degree Program

Supervised by:

Professor Michael Choi and Professor Simon Chesterman

AY 2022/2023

Yale-NUS College Capstone Project

<u>DECLARATION & CONSENT</u>

1. I declare that the product of this Project, the Thesis, is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic, or personal, in accordance with the academic regulations of Yale-NUS College.

2. I acknowledge that the Thesis is subject to the policies relating to Yale-NUS College Intellectual Property (<u>Yale-NUS HR 039</u>).

<u>ACCESS LEVEL</u>

3. I agree, in consultation with my supervisor(s), that the Thesis be given the access level specified below: [check one only]

☐ <u>Unrestricted access</u>
Make the Thesis immediately available for worldwide access.

☒ <u>Access restricted to Yale-NUS College for a limited period</u>
Make the Thesis immediately available for Yale-NUS College access only from <u>04/2023</u> (mm/yyyy) to <u>04/2025</u> (mm/yyyy), up to a maximum of 2 years for the following reason(s): (please specify; attach a separate sheet if necessary): <u>Intention to publish.</u>

After this period, the Thesis will be made available for worldwide access.

☐ <u>Other restrictions: (please specify if any part of your thesis should be restricted)</u>: NA

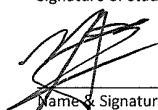<u>Tristan Koh Ly Wey (Cendana College)</u>
Name & Residential College of Student

_____
Signature of Student

3 April 2023
Date

Michael Choi
Name & Signature of Supervisor

3 April 2023
Date

Acknowledgements

Abstract

B.Sc (Hons) and L.L.B (Hons)

Ghosts in the Machine: Applying & Evaluating Explainable AI in Legal Decision Making

by Tristan KOH

This capstone assesses the explainability of the visualisations of machine learning models that were trained to identify data practices within apps' data privacy policies.

Natural language processing (NLP) techniques have recently increased in performance and can potentially automate some areas of substantive legal reasoning and writing, but has come at the cost of increased opacity of the models. Applying eXplainable AI (XAI) has been introduced to combat this opacity, but it is unclear whether such explanations are effective for the different purposes and users of the explanations. The stakeholders in data privacy regulation (i.e. organisations, data privacy regulatory authorities, and consumers of data services) can also benefit from the use of XAI. Therefore, there is a gap in research that assesses the effectiveness of XAI, and XAI's particular application in data privacy.

This capstone uses this data privacy context to train and evaluate AI classifiers that are able to detect data practices within PlayStore apps'

data privacy policies. XAI techniques are applied on these models to produce visualisations that aim to explain these predictions to these stakeholders in data privacy. Finally, the effectiveness of these visualisations are assessed by surveying the general public along differing metrics of explainability.

Word count: 11,226 (original = 9712 (incl footnotes), +50 for citations, and x1.15 for tables, content page, abstract, and acknowledgements.)

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke

Contents

1 Introduction

1.1 Motivation and significance

A question was posed in 2017: "Will there be a software service for the generation of explanations and arguments in law that will assist in structuring explanations of answers and supportive legal arguments?" [3]. This question seems close to getting an answer in 2023 with the release of ChatGPT [23] to the general public. Consider the following response from ChatGPT:
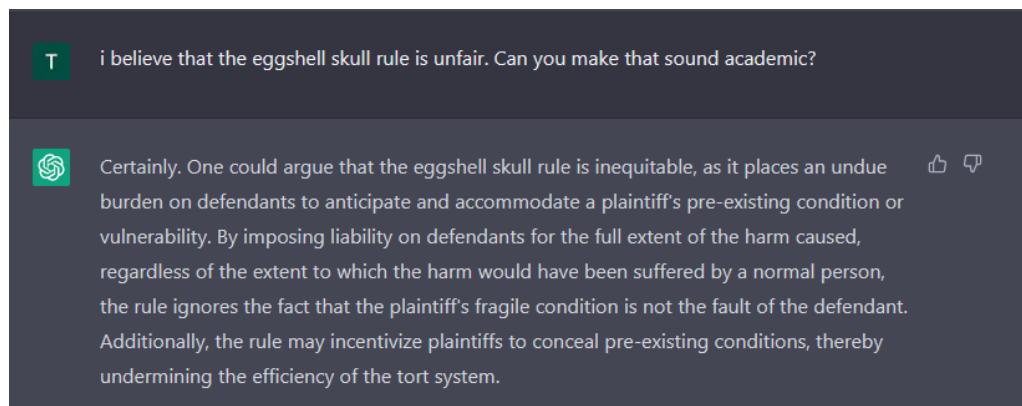
FIGURE 1.1: Conversation with ChatGPT about the eggshell skull rule.

Indeed, GPT-4, successor of ChatGPT and released a mere 4 months after ChatGPT, scored on the 90th percentile on the Uniform Bar Exam which is significantly in excess of the passing mark to practice law in the US [19].

ChatGPT and GPT-4 are the most recent AI models that uses advanced natural language processing (NLP). NLP is a branch of AI that gives computers the ability to understand text and spoken words as how human beings would understand [16]. While NLP techniques have substantially increased in performance in recent years, it has come at the cost of the explainability of their predictions because of models that are architecturally more complex [10]. This issue of explainability is exemplified by the last question I posed to ChatGPT about its view on the egg-shell skull rule:
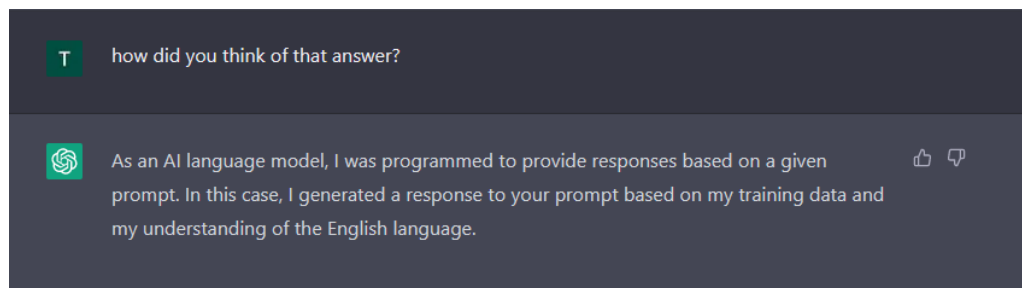
FIGURE 1.2: Conversation with ChatGPT about the egg-shell skull rule.

ChatGPT does not seem to be able to explain its views like how a typical human would[1]. This lack of explainability could be a significant hindrance towards NLP's greater adoption within the legal industry because the lawyer that uses these models ultimately bear the legal responsibility of ensuring that the analysis is legally sound. In Singapore, a legal practitioner must act with reasonable diligence and competence in the provision of services to the client[2]. A lawyer that relies on the analysis of legal tech tools but does not understand how the analysis was produced could be considered lacking in diligence and competence.

---

[1]Further prompting led ChatGPT to provide a list of academic papers that support its view. But there are issues of hallucination, where ChatGPT generates answers about facts that are false or non-existent [2].

[2]According to r5(2)(c) of the Legal Profession (Professional Conduct) Rules 2015.

Nevertheless, the intersection in skillset between data science and legal analysis is still nascent and it is unrealistic to expect all legally trained personnel to be trained in data science to the extent required to interpret the predictions of machine learning models without aid. Explainable AI (XAI) techniques and research have been rising in popularity since 2020 [20] but have not been adopted in legal tech. XAI refers to methods that attempt to help the end user of a model understand how it arrives at its result [8]. XAI for NLP has been designed to combat issues of opacity by reducing the abstraction of NLP techniques so that end users can understand how the model arrived at a decision. XAI can therefore bridge the gap between the lawyer and the data scientist by using XAI techniques to explain the predictions of legal tech AI models.

However, there is not much consensus about what "explainable" means. While the general agreed upon goal of XAI is to "completely, accurately and clearly quantify the [model's] logic", there is no consistent use of the terms "explainability", "interpretability" and "transparency" [8]. Interpretability is sometimes used to describe the model's internal logic and explainability as the ability of the user to understand that logic. Explainability is also sometimes described as techniques to explain the model's logic post-hoc without necessarily being representative of the model's true decision [27]. Since a debate on the "correct" definition of explainability is out of the scope of this capstone, I use explainability, interpretability, understandability and transparency interchangeably.

Another area of XAI is differentiating what explainability means to different users of the model who have different purposes for the explanations. For example, what could be explainable to data scientists may not

be explainable to laypersons. While data scientists may find that more technical details about the model would make the model more explainable, laypersons might be more confused if too many details are provided [27]. Therefore, assessing the effectiveness of XAI is highly dependent on the specific context and needs of the users.

In this capstone, XAI is assessed within the context of data privacy. The widespread collection and use of data by organisations in recent years has led to an increase of regulations governing data privacy, with 145 countries having enacted data protection legislation in 2021 [13]. With more sophisticated regulation comes increased difficulties for organisations to ensure that they are complying with these regulations, and for governments to enforce them. Given that explainability is context and user dependent, the below analysis explains why XAI is important to the three different stakeholders in data privacy: the consumer, organisation and state.

For the consumer, it is uncontroversial that data privacy policies on websites and software are rarely read, and even if they are read, consumers are unlikely to fully understand them because of the use of extensive legalese. The ease of reading data privacy policies was measured to be roughly the same as compared to academic articles such as the Harvard Law Review. The average policy takes 17 minutes to read, and the annual reading time per consumer is more than 400 hours which is more than an hour per day[3] [33]. This increasing unreadability of data privacy policies poses a significant challenge to the effectiveness of data privacy regulation given that the current model of regulation depends on the

---

[3]Assuming that a consumer visits 1462 unique websites each year.

consumer giving consent that is informed and freely given [22]. Combined with the increasing collection and use of data by organisations, consumers have diminishing control over how their data is collected and used. Therefore, XAI could be helpful to consumers since XAI can be used to automate legal analysis of data privacy policies and explain the implications of the analysis in simple terms.

Organisations in the EU are subjected to a data subject's "right to explanation" under the General Data Protection Regulation (GDPR). Data subjects have the right to obtain an explanation of a decision reached solely through automated processing[4]. For example, a bank could use AI to predict the probability of a customer defaulting on a loan. This prediction could be used to justify a decision to deny a loan to the customer. Under the GDPR, the customer has the right to not be subjected to such automated processing and obtain an explanation as to why the model made such a prediction. Therefore, the use of opaque AI could potentially affect the organisation meeting their legal obligations.

In Singapore, as part of Personal Data Protection Commission's (PDPC) guidance on the operations management of AI models, organisations are advised to provide explanations on how AI models are incorporated into the decision making process of the organisations so as to build understanding and trust with those stakeholders that use their products [25]. In terms of communications with their stakeholders, organisations are advised to develop policies on the type of explanations and when to provide them. Such communication could include explanations as to how the AI model was used in a specific decision.

---

[4]This is a plausible interpretation when reading Recital 71 in conjunction with Art 22 of the GDPR.

Despite debate about the scope of the right to explanation and whether it is binding [6], such legislation in addition to the PDPC's guidance for explainability supports the need for further development of XAI specifically in the context of data privacy. The models in this capstone can support the adoption of AI in organisations by providing explanations so as to aid the organisations in checking whether their privacy policies are in compliance with data privacy regulations.

One of the state's concerns when AI is used in legal decision making is the integrity of the judicial system and regulatory activities [7] since the legal system depends as much on the justification of its decisions as it does on the decisions themselves. In the context of Singapore's Personal Data Protection Act (PDPA), there is an appeal process for cases appearing before the PDPC and the higher courts can overturn the PDPC's decisions[5]. However, when overturning the PDPC's decisions, the higher courts do not disagree with the outcome but they disagree with the reasoning of the PDPC's decision. Assuming that AI models are sophisticated enough to write (or at least assist the writing) court judgements, if there was no explanation of the model's writing, there would be little basis for the higher courts to overturn judgements made by lower courts.

Hence, the importance of XAI applies across all three stakeholders in data privacy. Consumers can benefit from XAI by making privacy policies more readable. Organisations can use XAI to be more accountable to consumers and regulators by providing explanations for automated decisions. Transparency brought by XAI also helps the state to regulate organisations' use of AI.

---

[5]Per s48Q, s48R and s54 of the PDPA.

1.2   Research questions and roadmap

Therefore, I focus on NLP and XAI in the specific context of data privacy, as this context provides a realistic and practical scope for evaluating the explainability of AI models[6]. To this end, I have four research questions:

1. Which AI models ("classifiers")[7] perform the best on a data privacy dataset in terms of traditional performance metrics?

2. Which classifiers are the most explainable to users that include laypersons and users with domain knowledge in data science and the law?

3. How would users rate the explainability of a selected XAI technique?

4. What are the differences in explainability if users were asked to consider the predictions of these classifiers from the perspective of a consumer, an organisation or the PDPC?

To investigate these questions, I train and evaluate the performance of AI models that are able to classify sentences in apps' data privacy policies to a particular data practice. A privacy practice describes a certain behaviour of an app that can have privacy implications[8]. XAI methods are then used to visualise why the models made such predictions, and are assessed for effectiveness by conducting a survey asking respondents to rate these visualisations according to certain metrics related to explainability.

---

[6]All code and analysis can be found here.

[7]I refer to AI models as classifiers as the task that the AI models perform is classifying sentences to a data practice. I elaborate more on this distinction in Section 2.2.

[8]This will be further elaborated on when the dataset is introduced.

This capstone's objectives come broadly within the evaluation of XAI techniques. While there are studies that evaluate XAI and non-empirical research that investigates the relevant standard for explainability across different areas of law [14], I adopt an uniquely empirical methodology to assess XAI within a data privacy context across the different use cases of stakeholders. The closest study using similar methodology is one that asked lawyers to rate visualisations generated by different XAI methods [11]. However, this study was conducted using a dataset containing cases from the US Board of Veterans' Appeals, and did not consider evaluating explainability of different use cases.

Chapter 2 introduces the dataset and provides a literature review and justification for the methodology used in this capstone with regard to the NLP classifiers, XAI techniques, and methods used to assess XAI. Chapter 3 presents an exploratory data analysis of the dataset and reports the performance of the classifiers. Chapter 4 reports and discusses the results of the survey and Chapter 5 concludes the capstone with a summary of the findings in relation to the research questions, limitations and possible future work.

2 Literature review and methodology

There are four major components to this capstone: The dataset, model training, application of XAI techniques to the trained models and evaluating the effectiveness of these XAI techniques. I provide a literature review and describe the chosen methodology of these components below.

2.1 Dataset: The APP-350 Corpus

The APP-350 Corpus consists of 350 annotated Android app privacy policies. Each data practice ("practice") consists of a data type and a modality. A data type describes a certain behaviour of an app that can have privacy implications (e.g. collection of a phone's device identifier or sharing of its location with ad networks). There are two modalities: `PERFORMED` (i.e. a practice is explicitly described as being performed) and `NOT_PERFORMED` (i.e. a practice is explicitly described as not being performed). Altogether, 57 different practices were annotated. As not all practices had sufficient numbers to train models of sufficient baseline performance, I focused on

Data Type	Description
Contact_E_Mail_Address	Describes collection of the user's e-mail.
Identifier_Cookie_or_similar_Tech	Describes collection of the user's HTTP cookies, flash cookies, pixel tags, or similar identifiers.
Identifier_IP_Address	Describes collection of the user's IP address.
Location	Describes collection of the user's unspecified location data.

TABLE 2.1: List of top 4 data types and their descriptions.

Party	Description
1stParty	Describes collection of data from the user by the app publisher.
3rdParty	Describes collection of data from the user by ad networks, analytics services, or other third parties.

TABLE 2.2: List of modalities.

training models on the top 5 practices by frequency[1]. The data types and modalities that are used for the rest of the capstone are provided in Table 2.1 and 2.2.

The APP-350 Corpus was annotated and used in a broader project to train machine learning models to conduct a privacy census of 1,035,853 Android apps [37]. The authors downloaded the data privacy policies of apps from the Play Store with more than 350 million installs (which totalled 247 apps) and 103 randomly selected apps with 5 million installs. In total, there were data privacy policies of 350 apps. All 350 policies were annotated by one of the authors, a lawyer with experience in data privacy law. To ensure reliability of annotations, 2 other law students were hired to double annotate 10% of the corpus. With a mean of Krippendorff's $\alpha = 0.78$[2], the agreement between the annotations exceeded previous similar research.

---

[1]The performance of the models of the top n frequently occurring practices is later elaborated on in Chapter 4.

[2]Krippendorff's α is a measure of agreement, with $\alpha > 0.8$ indicating good agreement, $0.67 <= \alpha <= 0.8$ indicating fair agreement, and $\alpha < 0.67$ indicating doubtful agreement.

Since the focus of this capstone is to assess XAI methods within a data privacy context, the APP-350 corpus was chosen as it contains real-world data privacy practices from the PlayStore apps. Training models and evaluating XAI techniques on this dataset would provide a realistic indication of their performance. APP-350 is also a labelled dataset, allowing supervised training which provides a better validation of model performance.

The annotated privacy policies were originally in `.yml`, with one `.yml` file containing one app data privacy policy. Each data privacy policy is labelled at both the sentence and segment[3] level, but to limit the scope of this capstone I focus only on sentence level. The data was restructured from `.yml` to `.csv`.

2.2 Text representations and models

2.2.1 Literature review

In NLP, there are two components of a text classifier: The text representation, and the model. In this capstone, I distinguish between "model" and "classifier" for clarity. "Model" refers to the machine learning algorithm used to generate predictions, while "classifier" refers to the combination of a text representation and model that classifies a segment of text to a particular category. I make this distinction because part of this capstone's methodology is testing the explainability of the different combinations of models and text representations, of which both components are needed to form a classifier.

---

[3]A segment is similar to a paragraph.

Text representations are methods that translate plain text to mathematical representations that can be understood by a computer. The easiest and most intuitive text representation is a purely count-based approach such as Bag-of-Words, where each word is represented by their frequency in the sentence. These can be categorised as sparse vector representations. However, they do not capture the semantic and syntactic differences of words [21]. More advanced dense vector representations such as GloVe (Global Vectors) are better able to capture these differences. Generally, these dense vector representations are created through unsupervised learning. Neural networks are trained to predict the target word given surrounding words [21]. However, to capture such semantic meaning requires more abstraction and further increases the opacity and decreases the interpretability of NLP models. Hence, there is an inverse relationship between performance / opacity and interpretability. This is typically described as the "black-box" problem of AI: only the inputs and outputs to the system can be observed, but how the model derived the outputs from the inputs is not known (or at least not easily understood) because it is difficult to know exactly how the model is programmed [36]. In generating text representations of the APP-350 corpus, [37] used a union of Tf-IDF vectors and manually crafted features which were Boolean values indicating the presence or absence of frequently occurring keywords that occurred in each data practice.

Another component of a classifier is the model used to generate predictions. These models can be broadly categorised into classical machine learning models such as logistic regression and decision trees, and neural

networks such as GPT. However, while neural networks are higher performing, if simpler models are able to produce reasonable performance, these simpler models should be used instead. Much depends on the complexity of the text corpus, and the engineering task that the model are made for. Classical models are much less complex and therefore more explainable than neural networks, and require much less data to train. For the APP-350 corpus, [37] used support vector machines for classification (SVCs) to achieve reasonable performance. Individual models were then trained for every data practice. For all the classifications (except for four categories), they trained SVCs with a linear kernel and five-fold cross validation. For the four policy classifications, word-based rule classifiers were used instead because of the limited number of training data.

For the top 5 data practices that this capstone is concerned about, [37] were able to achieve F1 scores ranging from 77% (Identifier Cookie 1st Party), 91% (Contact Email Address 1st Party) to around 90% for location related data practices[4].

2.2.2 Methodology

Given that the focus is on explainability, and in any case the APP-350 corpus is probably too limited to train neural networks[5], I focus only on training classical models in this capstone. Since sparse vector representations and classical models have performed well for the APP-350 corpus

---

[4]Not all the data practices were used in training ($n = 188$) and testing ($n = 100$) these classifiers, so there is no reference performance for some of the data practices that are used in this capstone.

[5]For example, BERT was trained on the whole Wikipedia (and more), while GPT-2, the second iteration of GPT, was trained on 45 terabytes of data.

as seen above, I use SVCs and Tf-IDF. For the sake of comparison, I also use logistic regression and pre-trained GloVe embeddings.

The idea behind Tf-IDF is that words that are frequent in a document but rare across the rest of the corpus are more important in distinguishing the document from others. The Tf-IDF metric for a word in a document is calculated by multiplying two different metrics:

1. Term frequency (TF) of a word in a document. This is the number of times the word appears in a document.

2. Inverse document frequency (IDF) of the word across a set of documents. This is calculated by taking the total number of documents and dividing it by the number of documents that contain the specific word. This calculates the rarity of the term across all the documents. The closer the IDF of a word is to 0, the more common the word is.

In comparison, GloVe as a dense vector representation is able to better model semantic differences mathematically in the following way: $King - Man + Woman = Queen$ [32]. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The length-50 vectors used in this capstone are pre-trained using this algorithm on Wikipedia and news articles [24]. The main idea behind GloVe is to learn text representations by considering the statistics of words that occur together in a corpus. These statistics capture how frequently different pairs of words occur together in the text. GloVe builds a matrix from these statistics, where each element represents the number of times two words appear together in the same context. The matrix is then factorized into a product of

two low-rank matrices, where each row in one of the matrices represents a vector representation of a word. This results in dense, fixed-size vector representations for each word in the corpus that encode their meaning and relationships with other words in the corpus.

In terms of tokenisation, for both Tf-IDF and GloVe representations, text was lowercased and stop words were removed. No stemming and lemming was conducted. For Tf-IDF specifically, 1-gram to 4-grams were generated. For GloVe, length 50 pre-trained vectors were used to tokenise individual words, and then averaged across all of the words to represent one document.

In terms of models for prediction, I use SVC with a linear kernel. SVC works by finding the a linear hyperplane in a high-dimensional space that separates the different classes of data points with the maximum margin of separation (Figure 2.1). The hyperplane functions as a decision boundary and the margin is the distance between the hyperplane and the nearest datapoints from each class. I also use scikitlearn's SGDClassifier that implements linear SVC using gradient descent, which is another optimisation method apart from a linear kernel. Logistic regression is used as a baseline classifier for its simplicity. Logistic regression predicts the probability of a multi-class target variable as a function of the input features. The sigmoid function is usually used to transform the linear combination of features into a probability value between 0 and 1 (Figure 2.2). While I also trained ensemble classifiers (AdaBoost, GradientBoost, Random Forest), the performance of these models were roughly the same or less than logistic regression and SVC. Hence, I excluded them from further analysis.
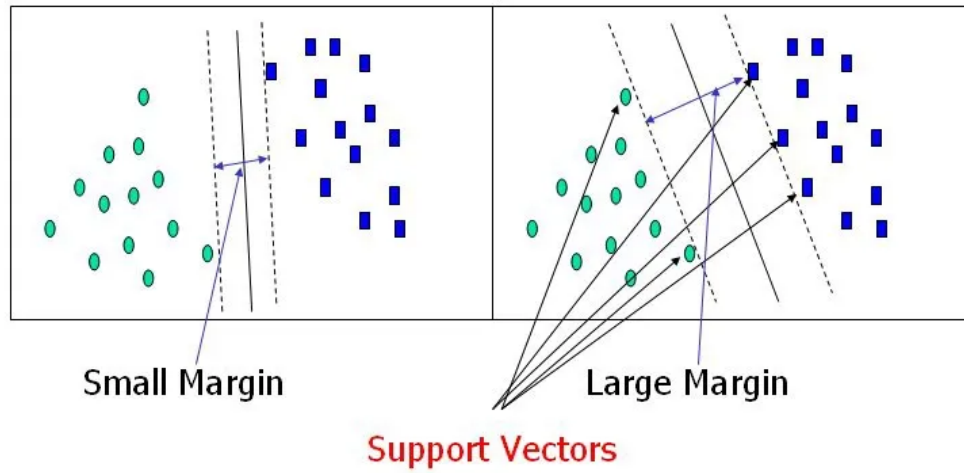
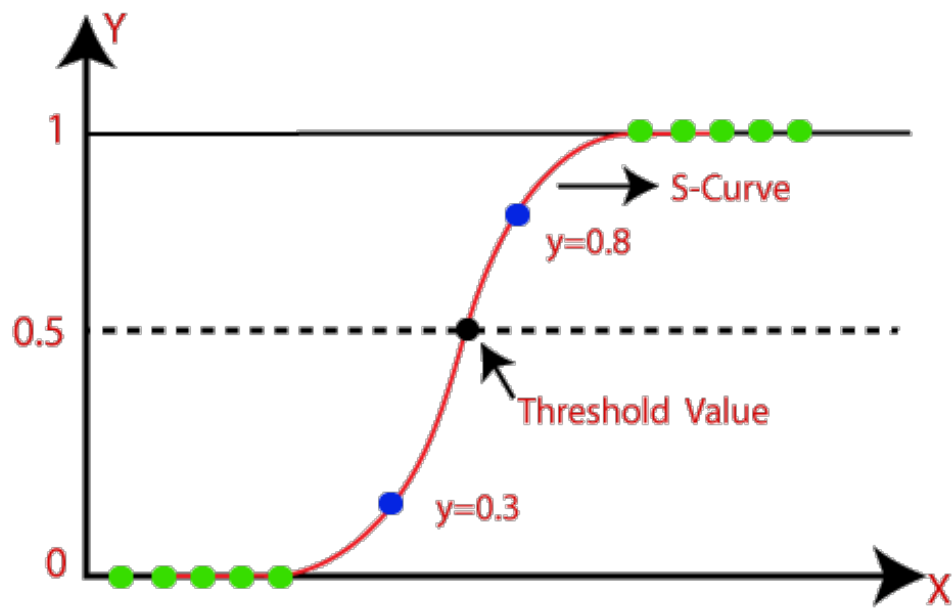FIGURE 2.1: Finding the hyperplane in SVC [4].

FIGURE 2.2: Sigmoid function in logistic regression [17].

2.3 XAI methods for NLP

2.3.1 Literature review

XAI for NLP have two broad characteristics: Local vs global, and self-explaining vs post-hoc. Thus, there are four possible categories of XAI as

seen in Table 2.3.

Category	Description
Local Post-Hoc	Explain a single prediction by performing additional operations (after the model has made a prediction)
Local Self-Explaining	Explain a single prediction using the model itself (calculated from information made available from the model as part of making the prediction)
Global Post-Hoc	Perform additional operations to explain the entire model's predictive reasoning
Global Self-Explaining	Use the predictive model itself to explain the entire model's predictive reasoning

TABLE 2.3: The four categories of XAI for NLP, adapted
from [8].

Some classical models such as logistic regression can be considered local and global self-explaining models. The coefficients of the features show the relative importance of each feature towards making the prediction. These coefficients are generated as part of the model training process and make it self-explaining. Logistic regression is also globally explainable since it models a deterministic function (sigmoid function), and this function serves as the "reason" for any prediction by the model. However, neural networks are not self-explaining by nature, in part as they are non-linear and are fitted onto high-dimensional data, in additional to the large numbers of parameters that make it difficult to ascertain the interaction between different features when a prediction is made. Therefore, post-hoc methods of explanability have to be used instead.

FIGURE 2.3: Sample visualisation of LIME

2.3.2 Methodology

I use LIME (Local Interpretable Model-agonistic Explanations), a local post-hoc XAI method that is also model agnostic [26]. Given a trained NLP model, LIME generates a set of perturbations on the input sentence by randomly replacing some of the words with similar words. These perturbed samples are passed to the NLP model for prediction. Using these predictions, LIME fits a simple self-explaining model (such as a linear regression or a decision tree) to explain these predictions. Through this local self-explaining model, LIME is able to generate the contribution of each feature to a particular prediction for a single sample. Figure 2.3 is an example of how an explanation of LIME can be visualised. Since LIME is a local, post-hoc method that uses a surrogate model to generate explanations, one disadvantage is that it may not fully represent how the underlying model actually made its prediction. Nevertheless, LIME has been found to be able to generate explanations that are faithful to the underlying model even just by using linear models for the surrogate model [26].

I chose to use LIME because it is model agnostic which allows easy experimentation with different combinations of text representations and models. Though logistic regression and SVC may be inherently self-explaining, I apply LIME on them because it visualises the feature importance without needing survey respondents to interpret coefficient scores of these models which can be inaccessible to those without a data science

background. I apply LIME on the different combinations of text representations and models and compare the explainability of LIME's explanations. I presented these explanations to the survey respondents without any amendments to the code output, except for minor editing to the labels in the figure because some labels were cut-off.

2.4   Evaluating the effectiveness of XAI methods

2.4.1   Literature review

There is no consensus of the methods or standards of assessing XAI techiques [31]. Standards of explainability are highly dependent on the purpose of generating the explanations. A few purposes have been suggested [9]:

1. *Global vs local:* A scientist inspecting a model predicting rainfall would require global explanation since he would be concerned with finding a long-term patterns, while a consumer who failed to obtain a loan would only require a local explanation to justify why the model made that specific decision.

2. *Area and severity of incompleteness of the problem:* A potential buyer of an autonomous vehicle (AV) would want a general overview of how AVs work, while a regulator may have specific explanation requirements such as whether a model detects a pedestrian 10 metres before collision.

3. *Nature of user expertise:* A layperson with no training in credit worthiness would not expect details of large complicated models in relation to the model's prediction of his credit rating, compared to an

auditor that would expect explanations of comparable granularity and nuance as human auditors.

Within the legal context, what could be insufficiently explainable for non-discrimination law [30] might not apply *mutatis mutandis* to the GDPR. Indeed, different standards for explainability have been argued to apply across contract, tort and banking law [14]. In any case, this capstone makes no claim as to the applicable standard for explainability for data privacy, and merely serves as empirical insight of explainability from the perspectives of the three stakeholders in data privacy.

In terms of methods of assessing XAI, human evaluation has been frequently utilised [8]. This involves asking humans to evaluate the effectiveness of the explanations in various ways (Table 2.4). The metrics being evaluated include fidelity (how much the explanation reflects the actual workings of the model), comprehensibility (how easily understood are the explanations), and others. However, many studies also do not state what is being evaluated.

Method	Description
Binary forced choice	Humans are presented with pairs of explanation, and must choose the one that they find of higher quality.
Forward simulation / prediction	Humans are presented with an explanation and an input, and must correctly simulate the model's output.
Counterfactual simulation	Humans are presented with an explanation, an input, and an output, and are asked what must be changed to change the model's prediction to a desired output.

TABLE 2.4: Methods of measuring human evaluation of XAI [9].

Human evaluation has been conducted within a legal context [11]. The authors trained a CNN on a dataset that contained 50 decisions issued by the Board of Veterans' Appeals of the US Department of Veterans

Affairs. Each sentence was annotated according to a type of legal reasoning (e.g. fact finding, legal reasoning, application of legal rule etc.). The focus of the research was to evaluate different XAI methods and lawyers' perceptions and expectations of XAI. Lawyers were asked to rate different XAI visualisations of the same 6 correctly classified sentences on a scale from 0 (worst) to 10 (best). Further, the lawyers that were surveyed were generally optimistic about the use of AI in law, with most agreeing that the explanations generated should be supplemented by further background knowledge, and that the model should be fair, lack bias, transparent and have awareness of the consequences of the AI-assisted / AI-automated decision. Lawyers also agreed that AI should increase automation and reduce their manual effort.

In addition, there is no agreed upon definition of "explainability". Explainability could be influenced by other value-laden terms, such as, *inter alia*, trust, effectiveness, fairness, bias, and the consequence of making a wrong prediction [27]. The relative importance of these values and overall requirement of explainability is necessarily dependent on the purpose and user of the explanation. Values such as fairness, bias, risk and trust could be argued to be of specific importance in a legal context which includes data privacy. The notion of "rule of law" encompasses these values and legal systems are arguably judged according to such values [12]. Nevertheless, there is no definitive position as to how these values can be used to evaluate XAI techniques that are specifically used within a legal context.

2.4.2 Methodology

With the foregoing discussion in mind, I expand upon the methodology of [11] by testing the effectiveness of LIME by varying the different purposes of explanations and methods of evaluation that were identified above. The primary method of evaluating LIME in this capstone is human evaluation through a survey. The survey was designed containing the following parts[6]:

1. **Part 1: Demographic data and beliefs relating to data privacy and AI.** Respondents were asked to provide their major or subject area expertise if they had graduated, and their level of experience with AI / data science and data privacy / law. This is used as a proxy for their level of expertise in AI or data science. Respondents are also asked to rate how much they think AI is useful, how far AI could be a risk, as well as how far they are concerned about their data privacy. This is used to understand the respondents' beliefs such that their assessment of LIME and the models can be evaluated in the context of these beliefs. Further cross-sectional analysis of the survey data can also be conducted using these data.

2. **Part 2: Rating of explainability metrics from the perspectives of three stakeholders in data privacy.** After a brief explanation of the data practices and how NLP works, respondents were provided with three contexts (app developer, PDPC & user). All three contexts relate to solely relying on the predictions of the classifier to

---

[6]The full survey can be found here, and a summary of the responses can be found here. Visualisations that were of significance to the analysis are provided later in Chapter 4.

make a decision in relation to a finding of a breach of the PDPA, and are meant to vary the purposes of the explanation according to the three purposes suggested by [9] above, as well as the legal & programming expertise of the end-user of the explanation. The consequence of making a wrong decision was kept constant across the contexts by stating that a $10,000 fine would be imposed as a consequence of a finding that the PDPA was breached.

After reading each context, the respondents were required to rate their beliefs relating to the use of the classifier in that particular context across four metrics: Effectiveness of model, Fairness, Risk to society, and trust in the model. These metrics were chosen as they are values of importance specifically to a legal decision making context. The purpose is to investigate how these values that relate to explainability can vary according to the purposes of the explanation.

3. **Part 3: Rating of LIME visualisations of four classifications by the same classifier (Logistic regression + Tf-IDF).** Respondents were asked to rate: (1) how far they understood why the model made the prediction, and (2) how far they found the visualisation easy to interpret. The intent is to evaluate both the underlying XAI technique (LIME) and the visualisation of the technique. For three classifications, respondents were given a counterfactual sentence replaced with words that were considered important by the classifier, and asked to predict whether the sentence would be classified under `Identifier Cookie 1st Party`. Counterfactual simulation is

meant to provide another method to measure explainability, aside from self-reported scores.

4. **Part 4 & Part 5: Choosing the more interpretable visualisation from a pair of identically classified sentences.** This uses binary forced choice to evaluate which text representation and model are more interpretable. Part 4 of the survey presents 3 pairs of visualisations from SVC + GloVe and Logistic regression + GloVe (controlling for text representation) and Part 5 presents another 3 pairs of visualisations from SVC + Tf-IDF and SVC + GloVe (controlling for model). Respondents were not told that the visualisations were produced by different models or text representations. The respondents' choices are then totalled up and compared against performance metrics (P/R/F1) of the classifiers to with these metrics acting as a "ground truth". This is to investigate whether there is a relation between interpretability and model performance.

5. **Part 6: Rating of the same explainability metrics for the same three contexts.** Respondents are finally asked to view the same three contexts and answer the same questions in Part 2, and their reported scores in Part 6 are tested for statistical increase / decrease compared to their initial scores in Part 2. This evaluates whether viewing the explanations significantly changed the respondents' beliefs of the use of the classifier in the three contexts according to the four metrics, and whether there are relations and trends between the individual metrics and contexts. Part 2 acts as a control variable in this comparison.

The survey was hosted on Google Forms, and respondents were anonymously recruited through friends and acquaintances of this author. The only restriction on recruitment was an age requirement of 18 and over for NUS / Yale-NUS students, or 21 and over for the general public due to Yale-NUS Ethics requirements. No personal identifying information of the respondents were collected.

3 EDA and classifier performance

3.1 EDA

There are a total of 18829 annotated sentences and the top 10 frequently occurring practices make up close to 60% of all sentences (Figure 3.1). The bottom 10 frequently occurring practices make up approximately 1% of the dataset. There does not seem to be much variation in sentence length for the top 10 frequently occurring practices, since the standard deviation for the mean is approximately 2.5 words and the median 1.9 words (Table 3.1) and as seen by similar interquartile range (IQR) (Figure 3.3c). This could indicate similar sentence complexity across the practices. The mean of the sentence lengths for all top 10 practices are slightly greater than the median, which suggest some outliers with very long sentences (Figure 3.2) that cause a slightly right skew distribution (Figure 3.3a). The distributions are also unimodal, with the mode around 20.

With regard to top 10 frequently occurring 4-grams in the top 5 practices (Figure 3.4), both 4-grams of `Cookie 1st Party` and `Cookie 3rd Party` contain many instances of "cookies", and it is hard to tell from the 4-grams alone whether they refer to collection of Cookies by 1st or 3rd parties. The same is seen to a lesser extent between the 4-grams of `Location` and `IP Address`, where both have instances of "IP address" and "mac address". In comparison, 4-grams of `Email` are clearly distinct from
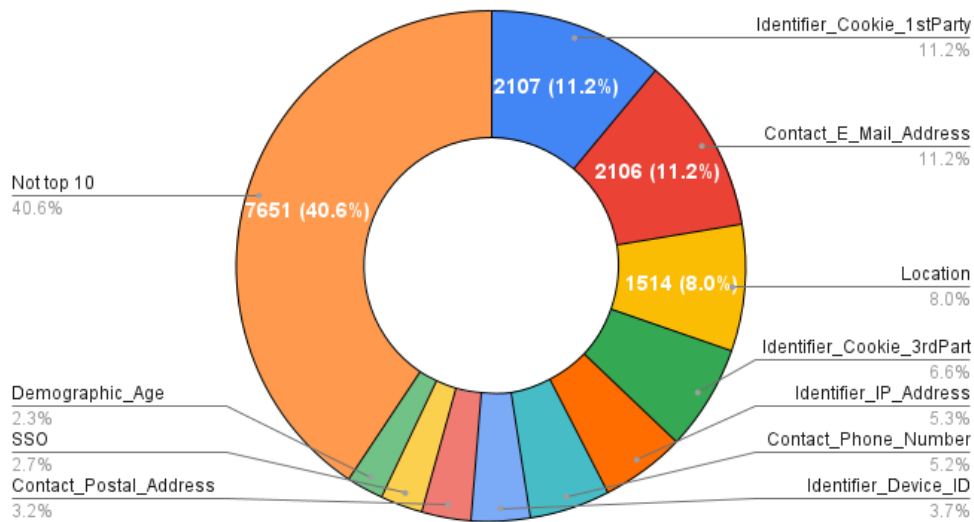
Data practices by proportion (n = 18,829)

FIGURE 3.1: Breakdown of data practices by proportion

	Mean sentence length	Median sentence length
count	10.0	10.0
mean	29.8	25.5
std	2.5	1.9
min	25.4	22.0
25%	28.7	24.3
50%	29.1	25.5
75%	32.0	26.8
max	33.1	28.0

TABLE 3.1: Summary statistics of top 10 occurring practices.

the rest of the data practices with unique tokens of "email" and "phone number".

3.2 Performance of classifiers for top N practices

I compare the weighted precision, recall and F1 scores of the classifiers below. Since there is neither a high risk of identifying false positives or
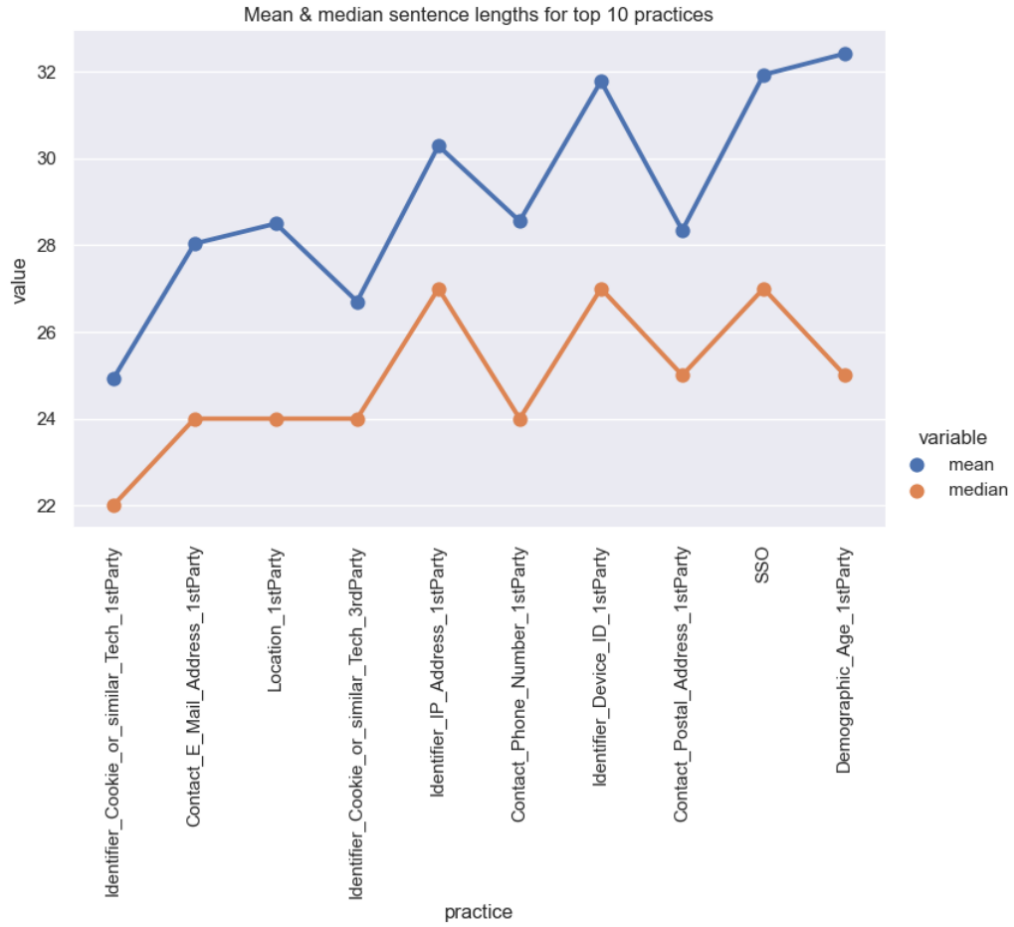
FIGURE 3.2: Mean and median sentence length

false negatives, F1 score is primarily used to assess the performance of the classifiers. As there is class imbalance in the top 10 data practices (Table 3.1), I focus on the weighted average F1 scores as the most important indicator of classifier performance.

Given that there are in total 57 practices for the entire dataset but the top 10 data practices already comprise 60% of all sentences, training classifiers on all 57 practices would likely lead to low performance. Thus, to find an optimal balance between classifier performance but still maintain realistic sentence complexity, I first assess the performance of the three

(A) Histogram of top 5 practices

(B) Histogram of next top 5 practices
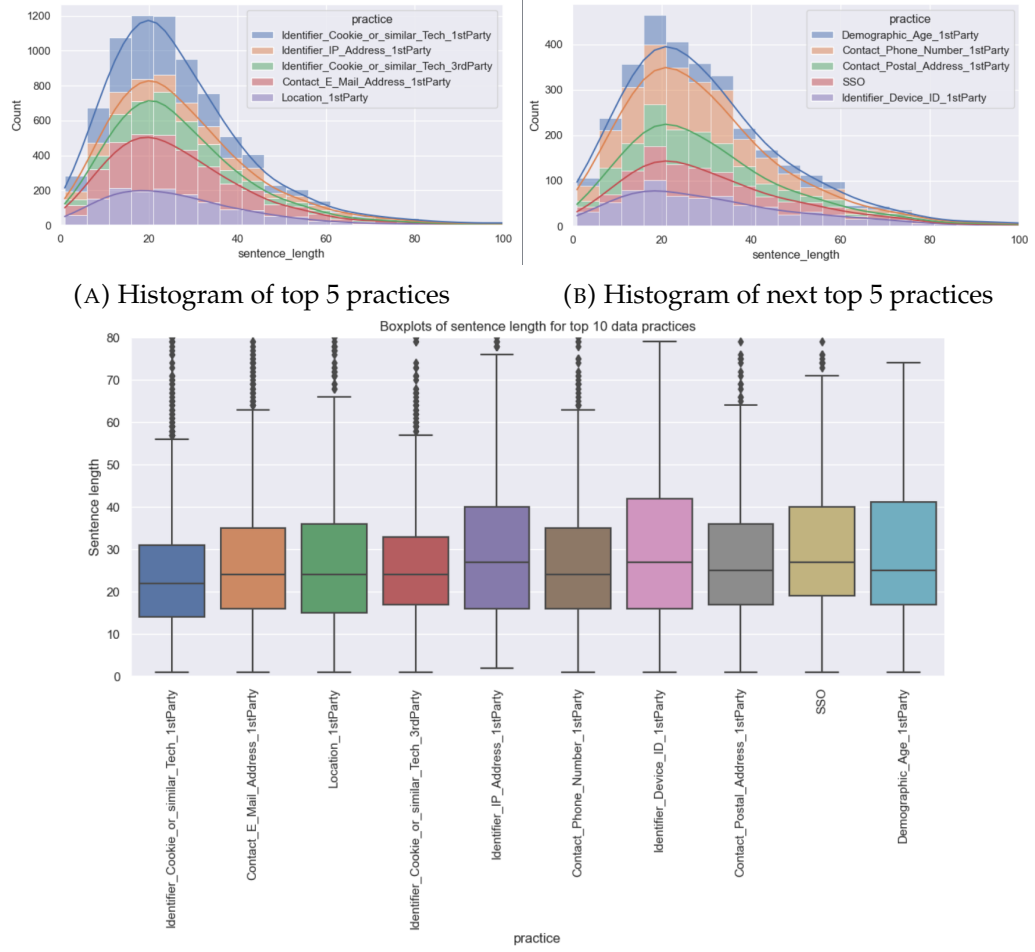
(C) Boxplots of sentence length (Practices shown in order of descending frequency)

FIGURE 3.3: Summary plots for top 10 frequently occurring data practices (note: not all outliers shown to focus on visualising IQR)

classifiers using Tf-IDF for the top N (where $3 \leq N \leq 10$) frequently occurring practices. SVC performs the best across the metrics and across the top N frequently occurring practices (Figure 3.5). This corresponds with the findings by [37]. Also, performance of all the classifiers generally decreases with increasing N, which is not surprising since it is more difficult for the classifier to find a decision boundary with more classes.
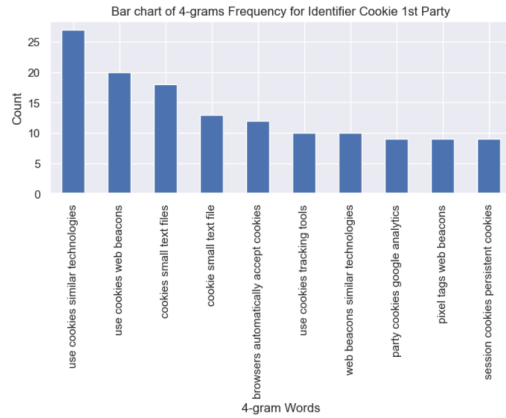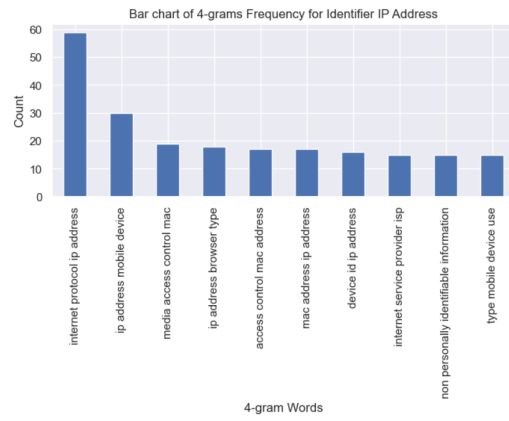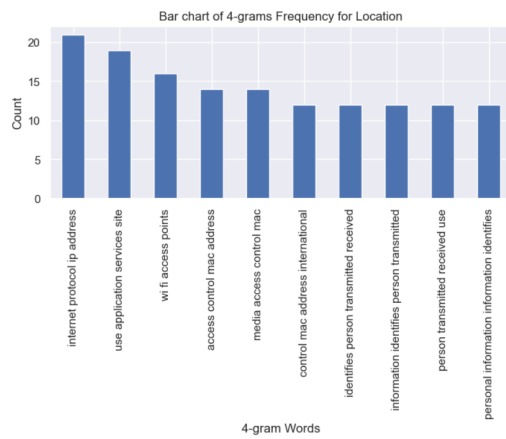
(A) Identifier Cookie 1st Party
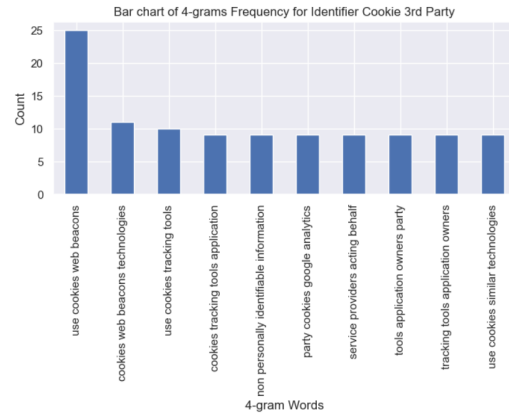
(B) Identifier Cookie 3rd Party
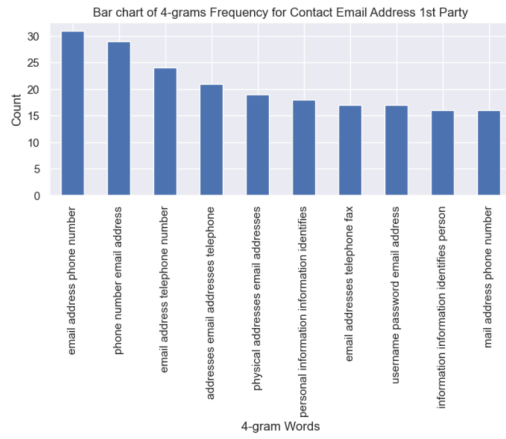
(C) Identifier Email

(D) Identifier IP Address

(E) Location 1st Party

FIGURE 3.4: Top 10 most frequently occurring 4-grams in the top most frequently occurring 5 data practices at the sentence level

The following sections focuses on classifier performance of the top 5 practices as performance was still reasonable at 60% weighted P/R/F1 scores. Also, since SGDClassifier performed the worst compared to the 2 other classifiers, I focus only on comparing logistic regression and SVC, with logistic regression as a baseline classifier.
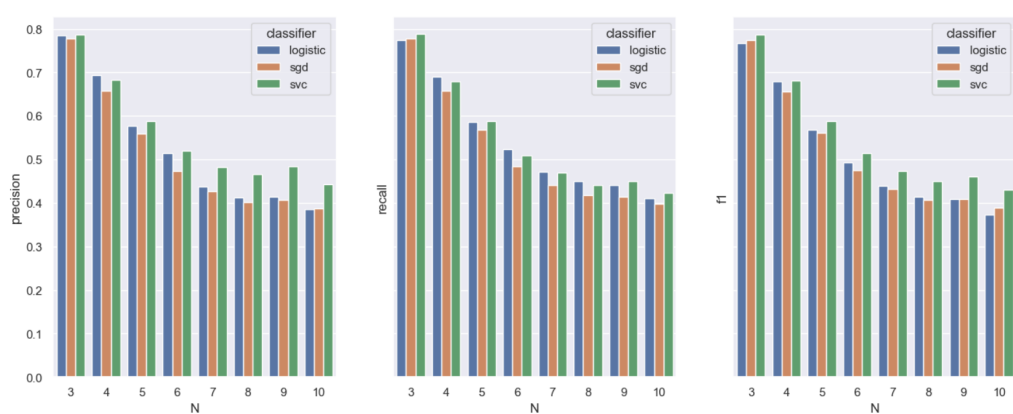
FIGURE 3.5: Weighted average P/R/F1 scores for top N
$(3 \leq N \leq 10)$ practices

3.3 Performance of individual classifiers for top 5 practices

I report the individual performance of all 4 combinations of text representations and models. With regard to P/R/F1 scores (Figure 3.6), similarly to what [37] found, SVC + Tf-IDF provided the best overall performance with the weighted average at 66%. This is quite a significant performance difference from the lowest performing model, logistic regression + GloVe at 54%. Specifically for `Cookie 1st Party`, both logistic regression + Tf-IDF and SVC + Tf-IDF performed similarly at 68% F1, but SVC + Tf-IDF

had a more balanced performance across precision and recall. With regard to the ROC curves (Figure 3.7), looking at the micro-averaged AUC, the classifiers all performed similarly well within a small difference of 0.03, and similarly for `Cookie 1st Party`, the difference in AUC was 0.02. GloVe embeddings seem to cause a greater variance when comparing the AUC for each class. For instance, the difference between the highest and lowest AUC for logistic regression + GloVe is $0.89 - 0.77 = 0.12$, while for logistic regression + Tf-IDF it was $0.91 - 0.85 = 0.06$.

Overall, these performance figures confirm that SVC is the better performing model, though the performance difference between the benchmark logistic regression model and SVC is not that significant (from 62% to 66% weighted average F1). It is also interesting to note that GloVe embeddings performed worse and caused higher variance in performance compared to Tf-IDF regardless of the model used, when GloVe is the more sophisticated text representation that can capture some semantic meaning of the tokens. These results correspond with a study measuring performance of supervised machine learning for classification tasks using generalised text datasets [15], where LinearSVC followed by logistic regression were the highest performing models.

(A) Logistic regression + TfIDF

(B) Logistic regression + GloVe

(C) SVC + TfIDF

(D) SVC + GloVe
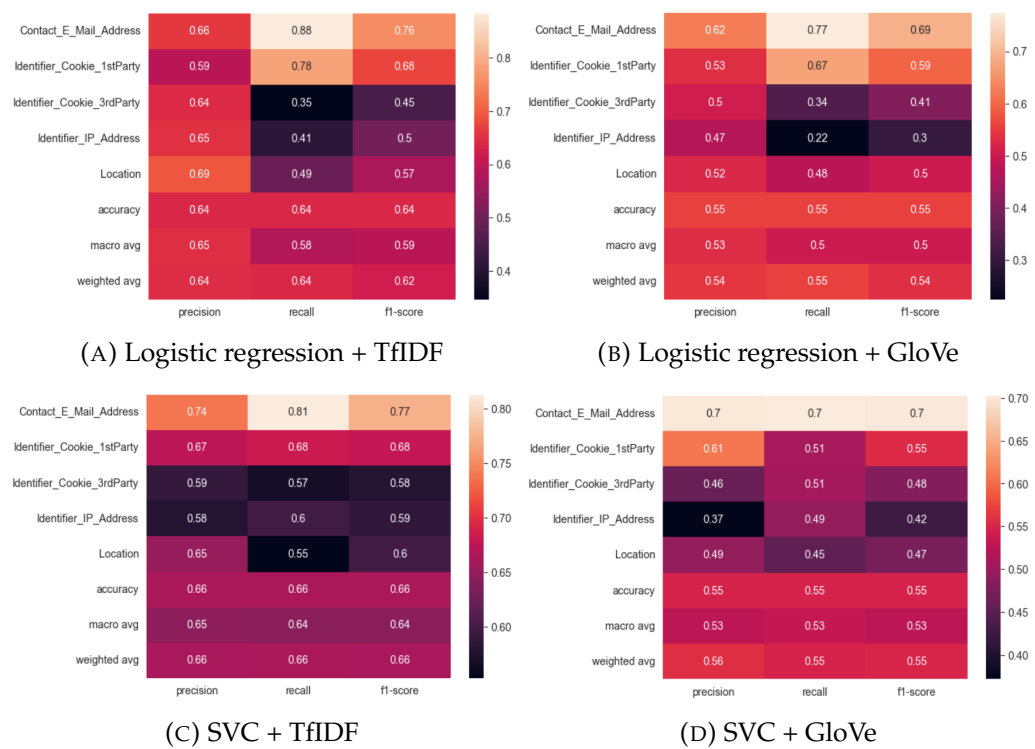
FIGURE 3.6: Performance heatmaps of the 4 classifiers (averaged over 5-fold cross validation)

(A) Logistic regression + TfIDF

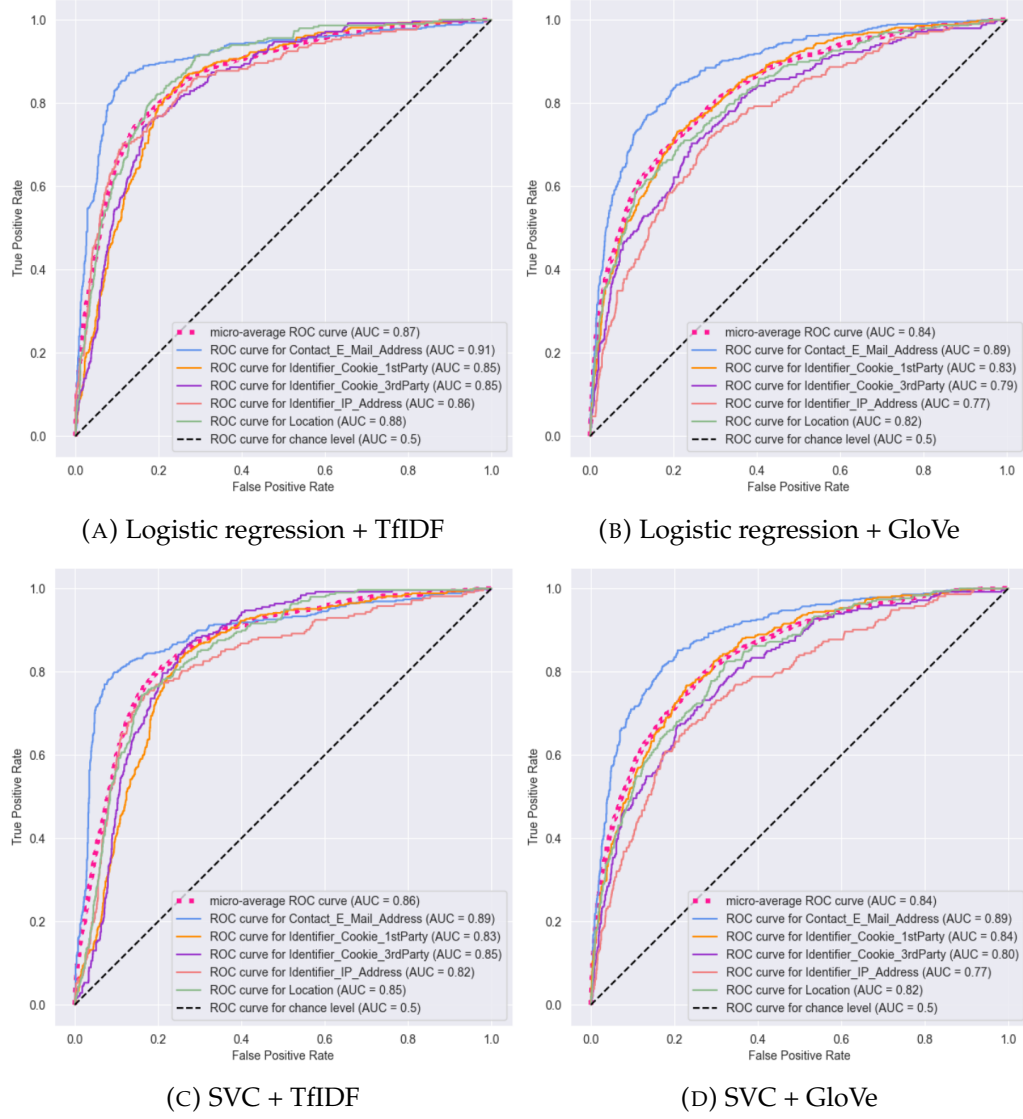(B) Logistic regression + GloVe

(C) SVC + TfIDF

(D) SVC + GloVe

FIGURE 3.7: Multiclass Receiver Operating Characteristic
(ROC) Curve of the 5 data practices for each classifier

4 Discussion of survey results

In total, 31 responses were collected. The majority of respondents majored in law (58%) vs non-law (42%) respondents (Figure 4.1). Except for the questions relating to subject matter expertise (data privacy and AI), the level of agreement of law vs non-law respondents about their beliefs relating to data privacy and AI were about the same (Figure 4.2). Law respondents had less expertise in AI, while conversely, non-law respondents had less experience with data privacy. Across all respondents, while they rated that decisions by AI could be a risk to society (about 4), they also agreed that decisions by AI could be equally useful. This perhaps suggests that the respondents think the balance between "usefulness" and "risks" are not zero-sum; AI could be very helpful in solving problems, but at the same time users should be cognisant of the risks.

I mainly used sentences annotated as `Identifier Cookie 1st Party` to produce explanations for the survey. This is because the class performance of `Identifier Cookie 1st Party` performed relatively well, and has the highest number of instances in the dataset which provides a large variety of drafting styles.
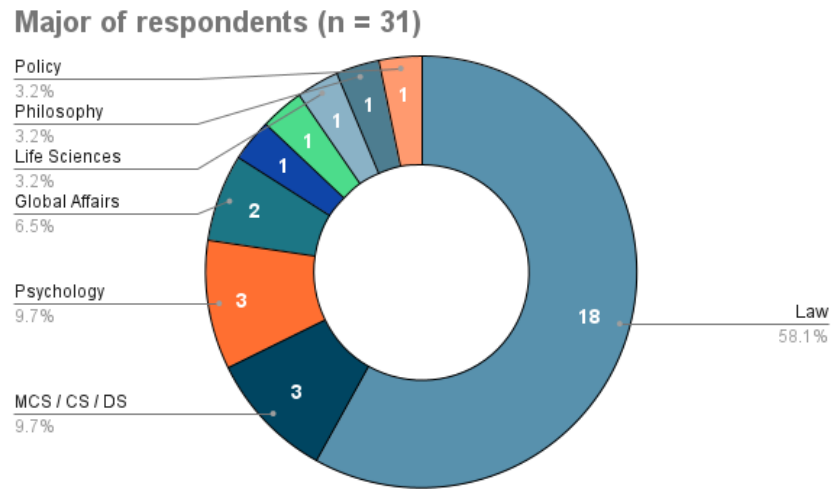
FIGURE 4.1: Breakdown of respondents' expertise by major

4.1 Part 2 & Part 6: Comparison of self-reported scores of explainability across the three contexts

Using the Wilcoxon Rank Sum Test, I tested for the following, setting $\alpha = 0.1$:

$H0$: There is no increase / decrease in scores after viewing the explanations.

$H1$: There is an increase / decrease in scores after viewing the explanations.

The 1-sided test was used to check whether the distribution underlying the difference between the initial and final paired scores was symmetric below or above 0 [28]. Mathematically this difference can be stated as $d = i - f$, where i and f are the scores reported before and after viewing the explanations, and d is the difference. Hence, if $d < 0$, then $i < f$ and the scores increased after viewing. Conversely, if $d > 0$, then $i > f$
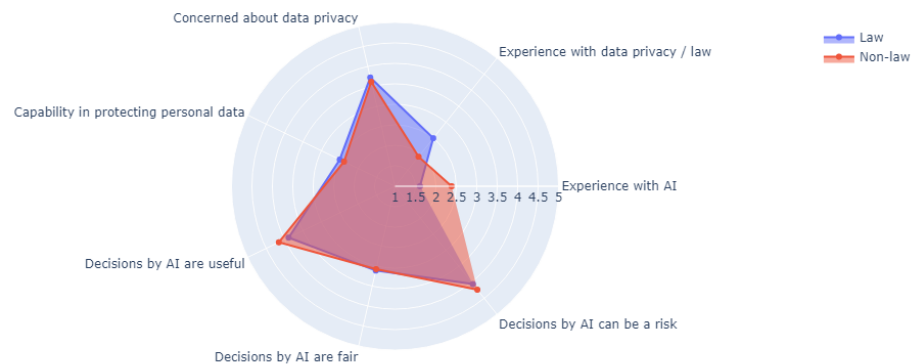
(A) Law vs Non-law respondents
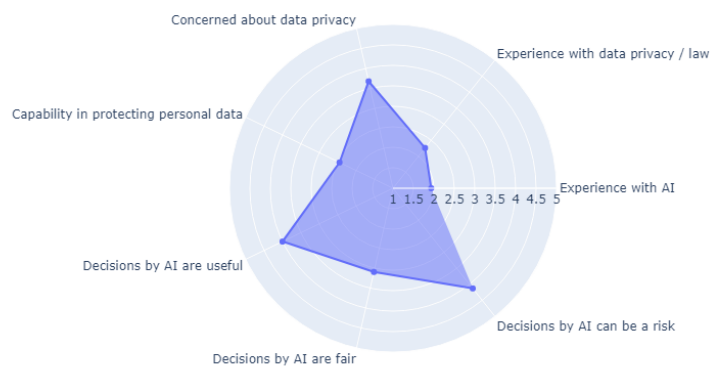
(B) All respondents

FIGURE 4.2: Mean scores of self-reported beliefs of respondents regarding AI & data privacy. (1 = least agree, 5 = strongly agree. $n = 31$)

and the scores decreased after viewing. p-values are reported in Table 4.1 and 4.2. For Table 4.2a and 4.2b, I took the mean of self-reported scores across contexts, and across each metric respectively. For Table 4.2c, the self-reported scores were averaged across both context and metric. The trending in scores across the contexts are visualised in Figure 4.3.

Question	Context 1: Increase	Context 1: Decrease	Context 2: Increase	Context 2: Decrease	Context 3: Increase	Context 3: Decrease
Do you think model is effective?	0.013	0.987	0.932	0.0684	0.856	0.144
Do you think model is a fair method?	0.382	0.618	0.841	0.159	0.933	0.0671
Do you think model is a risk to society?	0.756	0.244	0.428	0.572	0.825	0.175
Do you trust the prediction of the model?	0.887	0.113	0.945	0.055	0.837	0.163

TABLE 4.1: p-values comparing whether there was a statistically significant increase / decrease in the explainability scores before and after viewing explanations.

Context	p-value
1: Increase	0.369
1: Decrease	0.633
2: Increase	0.940
2: Decrease	0.060
3: Increase	0.955
3: Decrease	0.0450

(A) p-values by metrics, taking the mean of scores across contexts

Metric	p-value
Effective: Increase	0.485
Effective: Decrease	0.515
Fair: Increase	0.826
Fair: Decrease	0.174
Risk: Increase	0.660
Risk : Decrease	0.340
Trust: Increase	0.953
Trust: Decrease	0.0468

(B) p-values by contexts, taking the mean of scores across metrics

Increase	Decrease
0.844	0.156

(C) p-values taking the mean scores across contexts and metrics

TABLE 4.2: p-values comparing mean scores (cells highlighted in red are statistically significant)

Here are the instances when $p < 0.1$, and therefore $H0$ can be rejected in favour of $H1$:

1. Effectiveness and trust significantly decreased for context 2 (PDPC), and fairness significantly decreased for context 3 (user). Effectiveness significantly increased for context 1 (app developer) (Table 4.1).

2. There is a statistically significant decrease in explainability metrics for context 1 and 2 when comparing the mean of the scores across
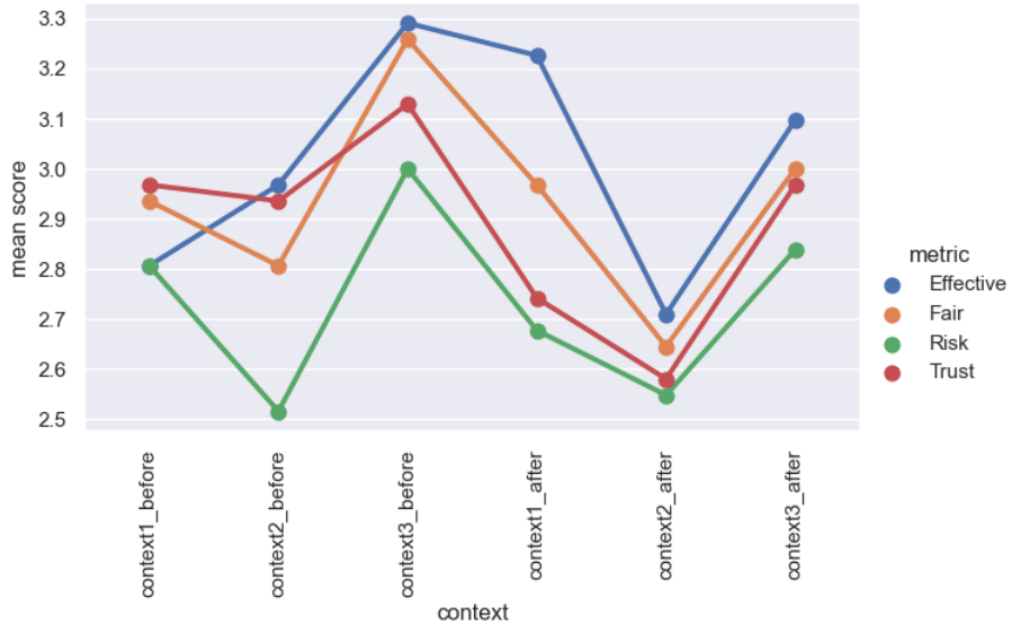
FIGURE 4.3: Mean scores for each metric and each context
before and after respondents viewed explanations

the metrics (Table 4.2a) and for trust across the three contexts (Table 4.2b). Interestingly, there seems to be a higher correlation between the scores of the 4 metrics specifically for context 2 and 3 after the the respondents viewed the explanations (Figure 4.3).

Remember that the explanations given to respondents were deliberately chosen to demonstrate the limitations of the model. Hence, respondents were likely more cognisant of such limitations[1] when they answered the same questions again in Part 6. Here are some inferences that I draw from these observations:

1. *"Efficiency" is dependent on the risks of making wrong predictions (App developer, PDPC)*: The speed of automation of reading privacy policies would be the same regardless whether an app developer or

---

[1]This can also be inferred from the negative trend of interpretability and understandability as explained below.

PDPC uses it. However, changes in efficiency differed between the 2 contexts, suggesting that automation was not the only metric that respondents considered as part of "effectiveness". Respondents could have considered that automation was more important to app developers when balanced against the risks of making a wrong prediction, as compared to the PDPC, where the risks were too high to justify automation. Hence, the classifier is not "effective" in terms of making a legally defensible decision given that it is fallible.

2. *Classifier's interpretability is especially critical for the PDPC such that trust was affected*: Respondents could have thought that PDPC, as a regulatory body in comparison to app developers and users who are not, should fully understand and explicitly justify the decision of finding an organisation (such as an app developer) liable for breaching the PDPA. Hence, when respondents reported a drop in overall interpretability and understandability, it especially affected their trust in the logic of the classifier.

3. *Fairness is perceived in relation to the user's own interests, rather than systematic fairness*: If respondents believed that the logic of the classifier was unfair because it would unfairly discriminate against a certain style of writing data privacy policies (a kind of "systemic fairness"), respondents would also have reported a drop in fairness for the PDPC context as there would be the most expectation on a regulatory body to treat data subjects and organisations equally. However, fairness only dropped for the user and not the PDPC. This could suggest that respondents viewed the drop in fairness as not being able to get recourse for their data rights which have been

violated because the classifier they are relying on for evidence to the PDPC could make wrong predictions which would affect their legal claim.

4. *Expertise of the end user of explanations affects explainability*: In relation to the overall drop in explainability metrics for the contexts of the PDPC and the user, perhaps the respondents thought that PDPC and the user did not have technical knowledge to better understand the classifier's logic beyond the explanations that were presented to them, whereas the app developer would have such technical knowledge. Hence, the lack of technical expertise of the PDPC and the user caused the drop in explainability metrics.

5. *Mismatch in expectations particularly affects trust*: In the case of trust being the only metric that significantly decreased across the contexts, this could point to a mismatch of expectations. Intuitively, if respondents do not understand how a process works, they are less likely to trust the results of that process. Perhaps respondents' expectations of AI was that AI would reason similarly to how humans would, or that AI functioned entirely objectively similar to applying a formula in Excel where it is clear how the result was obtained from the formula. Instead, the explanations provided showed that the classifier was inconsistent, sometimes relying heavily on words like "cookies" and other times not so much without any discernible reason. So respondents' expectations were not met, and this caused the overall drop in trust.

4.2 Part 3: Testing whether viewing more visualisations increased explainability

4.2.1 Analysis of the reported scores of interpret and understand

There is a decreasing trend of both understandability and interpretability after viewing each explanation (Figure 4.4, explanations can be viewed in Figure 4.5). Using the Wilcoxon Rank Sum Test, I separately tested for significant differences of understandability and interpretability, setting $\alpha = 0.1$:

H0: There is no difference in reported interpretability / understandability between the first and last questions.

H1: There is an increase / decrease of reported interpretability / understandability between the first and last questions.

	Increase	Decrease
Interpretable	0.867	0.132
Understandable	0.999	<0.001

TABLE 4.3: p-values comparing reported understandability and interpretability between the first and last question (cells highlighted in red are statistically significant)

There was a statistically significant decrease in understandability but not interpretability (Table 4.3). Hence, H0 can be rejected in favour of H1.

A reason for why there was no significant decrease in interpretability was because the questions I posed to the respondents were clear in distinguishing the visualisation from the classifier's logic. Hence, respondents understood how to interpret the visualisation, but were less clear about the classifier's logic that was presented through the visualisation.

Mean scores of understandability and interpretability after viewing each explanation
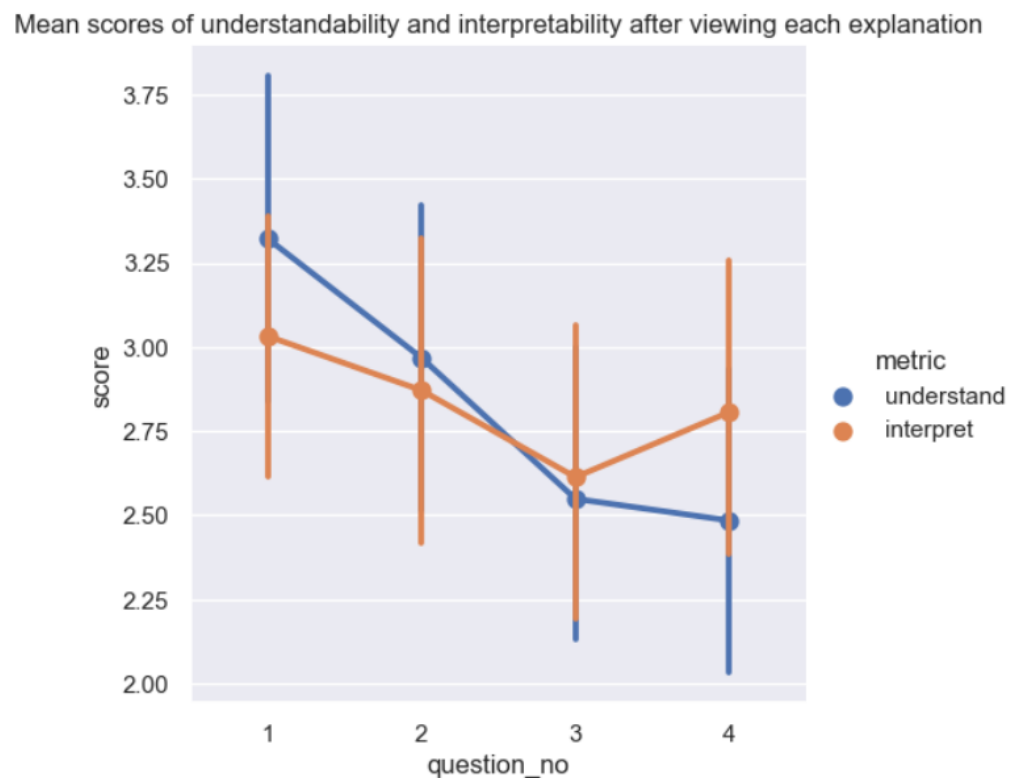

FIGURE 4.4: Trend of the mean of self-reported understanding and interpretability after viewing each explanation

With regards to the significant decrease in understandability, the respondents could be confused about how the model works globally after being exposed to specific predictions that were in themselves explainable, rather than being confused about a specific prediction of the model. As mentioned, the visualisations for this part were specifically chosen to demonstrate the limits of the classifier by changing keywords which were strong predictors of the data practice. This means that respondents were being exposed to a more complex (and confusing) of the model globally as they found contradictions in how the model used certain keywords in some examples but not in others. Therefore, each explanation was actually effective in communicating to the respondents how that particular
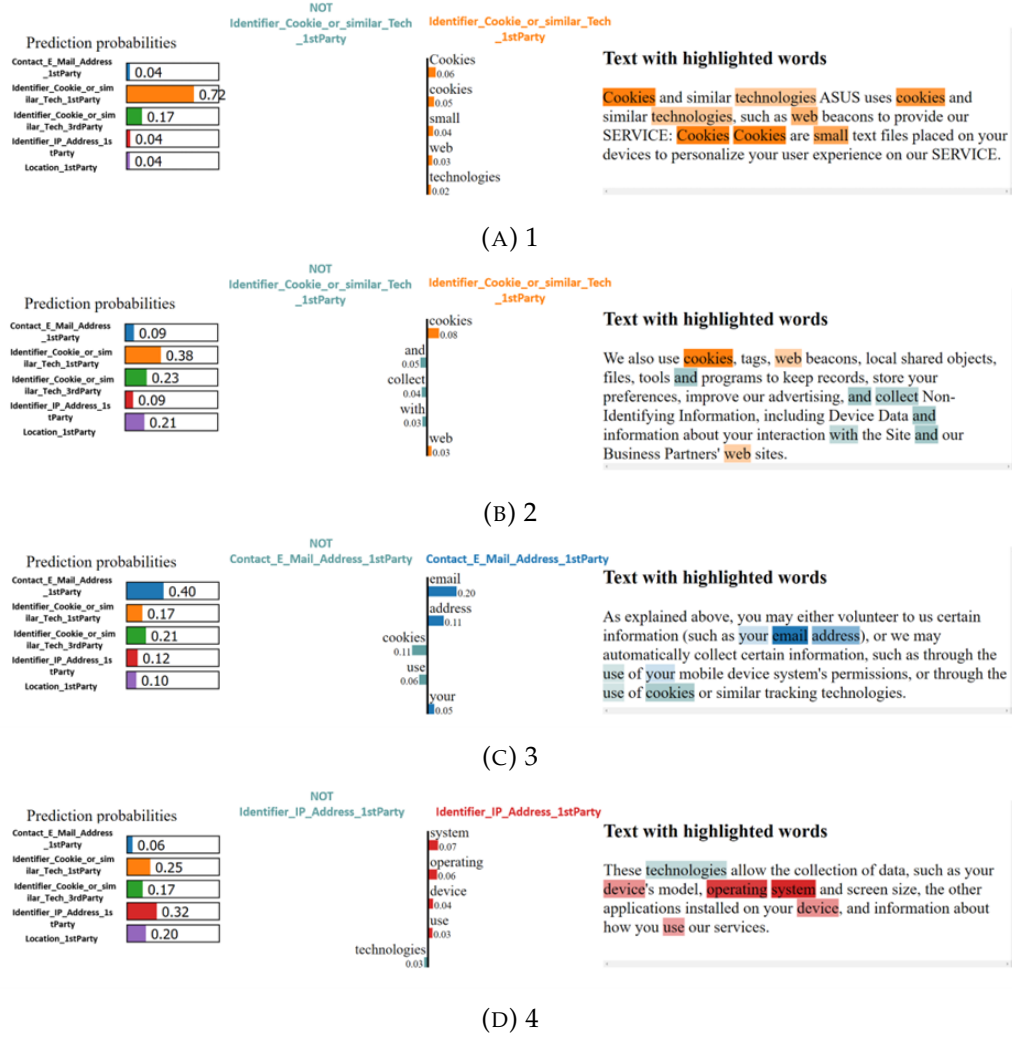
(A) 1

(B) 2

(C) 3

(D) 4

FIGURE 4.5: The 4 visualisations shown to respondents. Respondents were asked to rate how far they understood why the model made the prediction, and how far they found the visualisation easy to interpret. All 4 sentences were annotated as `Identifier Cookie 1st Party`.

prediction was made, and therefore could be considered as locally explainable. However, as a whole, respondents' understandability of the model given the contradictions seen through comparing each explanation decreased. Therefore, the classifier could be said to be locally understandable, but globally less understandable.

Another inference is that the underlying explainability method also

influences the respondents' perception of the interpretability of the visualisation technique (or vice versa). This can be seen by how there is a positive correlation between the understand and interpret scores. This is not surprising since if the visualisation technique is unclear or poor, understandability of the model itself would also decrease since respondents' only way of viewing the results of the explainability technique is through the visualisation technique. To respondents, both the visualisation technique and the explainability technique are one and the same thing.

4.2.2 Analysis of predicting counterfactuals

Given the lack of any consistent trend of the respondents' votes across the questions (Figure 4.7), it is difficult to draw any definitive interpretations from the respondents' predictions of counterfactuals. Further, the classifier itself gave ambivalent predicted probabilities for each data practice. For example (Figure 4.6b), the classifier predicted `Contact Email Address` at 25% while `Cookie 1st Party` was predicted at 23%.

4.3 Part 4 & 5: Testing which model and text representation is more explainable

As there were three questions in each section to test the explainability of each pair of text representation and model, I totalled up the votes for each option for each part and each question (Figure 4.8 and 4.9).

Overall, respondents found no difference in understandability between logistic regression and SVC (Figure 4.8), while it was more contentious

(A) Original explanation
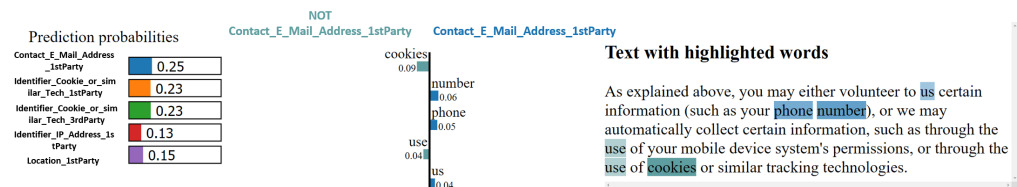
(B) Counterfactual explanation (not shown to respondents)

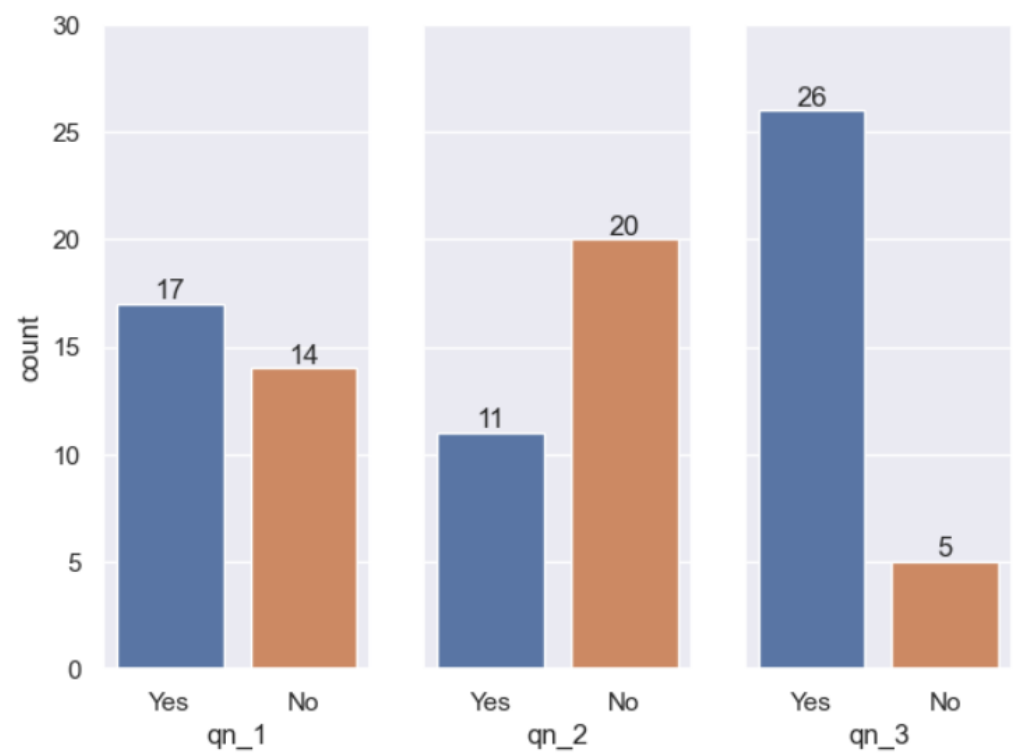FIGURE 4.6: Sample counterfactual explanation (1 out of 3)

FIGURE 4.7: Votes for predicting whether counterfactual
would be classified as `Identifier_Cookie_1st_Party`

when comparing text representations, with a third split across the three
options (Figure 4.9). Remember that classifier performance for each pair
actually did not differ significantly. In terms of comparing respondents'

votes with the "ground truth" metric of model performance, most respondents would be expected to indicate that there were minimal differences between the understandability of the classifiers. This was seen when comparing logistic regression vs SVC but not seen when comparing Tf-IDF and GloVe. While there is no majority consensus for the comparison of text representations, with votes which were almost equally split instead of being heavily weighted in favour of one option show that respondents were more undecided in the understandability of the models. One inference of such results is that while respondents collectively did not give a definitive answer as to the more explainable word representation, there are still possible differences in understandability even when comparing classifiers of relatively similar performance.

(A) Total votes across 3 questions
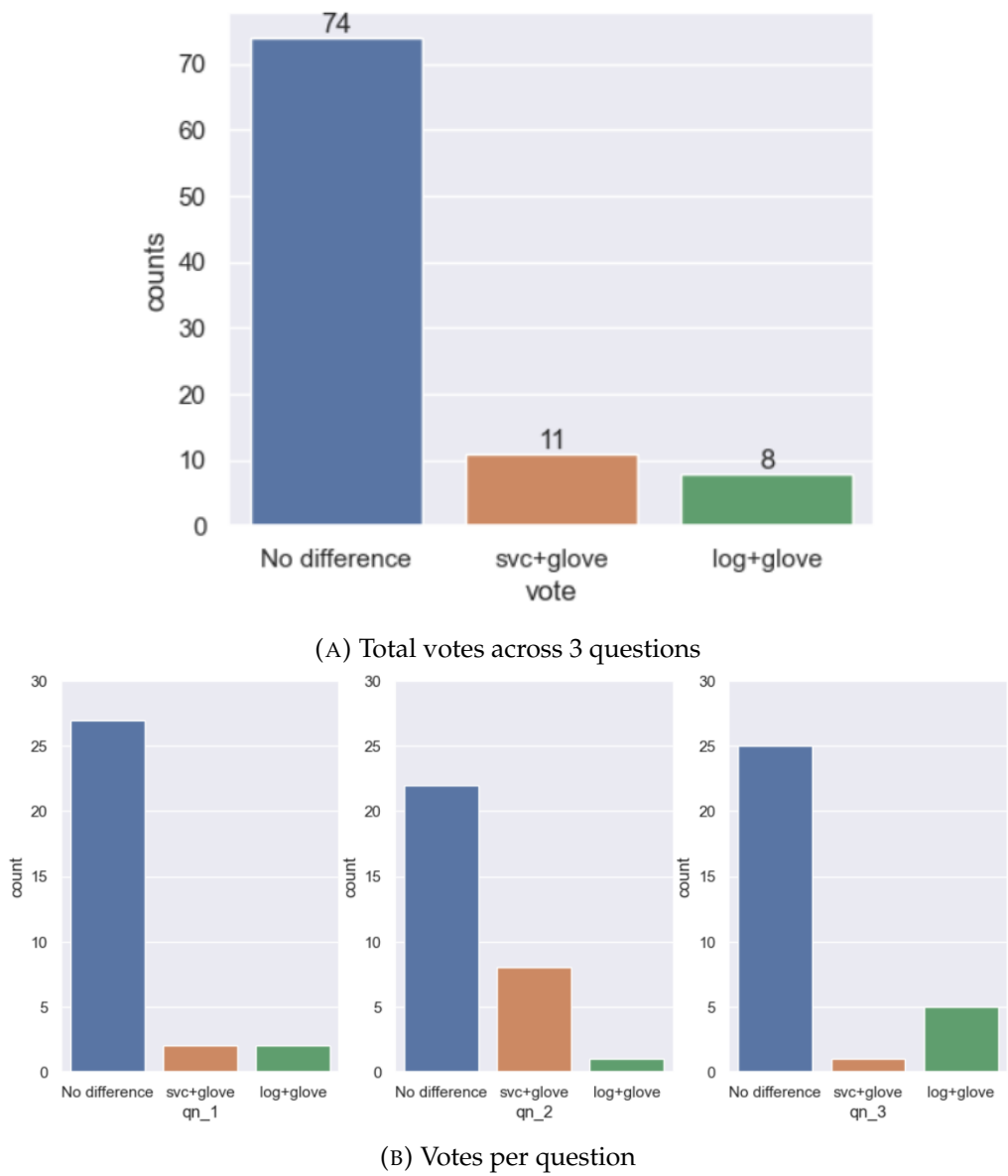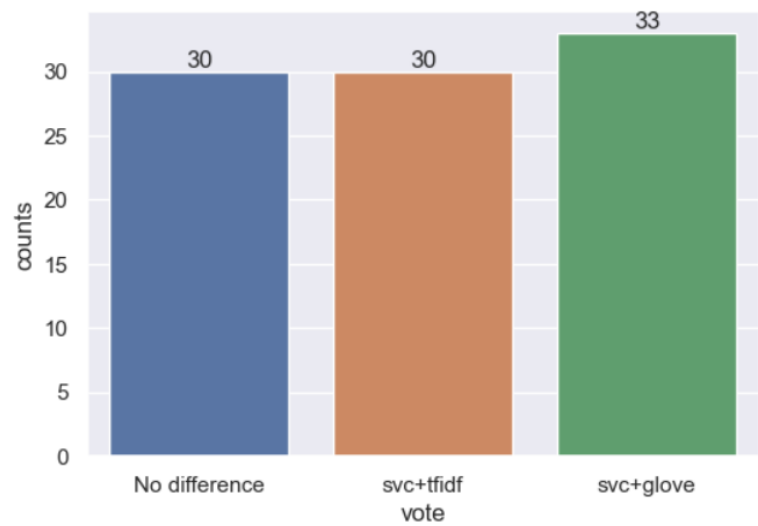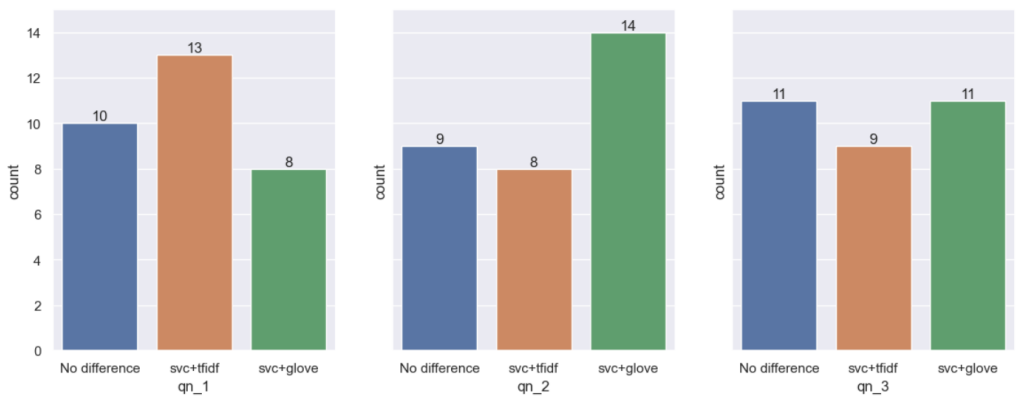

(B) Votes per question

FIGURE 4.8: Respondents' votes to whether SVC + GloVe
or Logistic regression + GloVe were more understandable

(A) Total votes across 3 questions

(B) Votes per question

FIGURE 4.9: Respondents' votes to whether SVC + TfIDF
or SVC + GloVe were more understandable

5 Conclusion

In summary, this capstone has conducted an uniquely empirical evaluation of the XAI of NLP within a data privacy context from two perspectives: The first through traditional performance metrics of machine learning, and the second through human evaluation. The survey investigated how explainability interacted with values that are particularly important to a legal domain, and tested the effectiveness of LIME through different types of questions posed to respondents such as counterfactual reasoning.

5.1 Findings

The research questions that were posed in Section 1.2 can now be answered:

1. **Which classifiers perform the best on a data privacy dataset in terms of traditional performance metrics?**

   Overall, SVC + Tf-IDF performed the best when measuring performance using weighted F1, though there was not much difference when comparing micro-averaged AUC for the ROC curves. In terms of performance when classifying `Identifier Cookie 1st Party`, both SVC + Tf-IDF and logistic regression + Tf-IDF performed similarly. It is surprising that GloVe embeddings performed worse

overall and caused greater variance in the AUC of all the classes compared to Tf-IDF, given that GloVe accounts for semantic context as compared to Tf-IDF which is purely count-based.

2. **Which classifiers are the most understandable to users that include laypersons and users with domain knowledge in data science and the law?**

   When SVC was compared to logistic regression, respondents by a large majority found no difference in the understandability. When Tf-IDF was compared to GloVe, respondents were undecided across the three options. Comparing the respondents' votes with performance metrics, there seems to be a slight (but non-statistical) relationship between understandability and classifier performance. Tf-IDF consistently outperformed GloVe regardless of the model used, which is reflected in how respondents were equally split between the 3 options. Whereas comparing logistic regression and SVC, the latter did not consistently outperform the former, and majority of respondents similarly found no difference between the explanations.

3. **How would users rate the explainability of a selected XAI technique?**

   There was a statistically significant decrease in respondents' reported understandability but not interpretability of LIME, and there was a negative trend of both understandability and interpretability across

the questions. One interpretation is that global rather than local explainability decreased because respondents were exposed to contradictory information about the model with each new explanation they viewed.

4. **What are the differences in explainability if users were asked to consider the predictions of these classifiers from the perspective of a consumer, an organisation or the PDPC?**

Effectiveness and trust significantly decreased for the PDPC context, and fairness for the consumer context. Only effectiveness significantly increased for the app developer context. Explainability metrics significantly decreased for PDPC and consumer context, and overall trust also significantly decreased. Some inferences of these results were:

(a) Efficiency is dependent on the risks of making wrong predictions,

(b) Classifiers' interpretability is especially critical for the PDPC such that trust was affected,

(c) Fairness is perceived in relation to the user's own interests as opposed to systemic fairness,

(d) Expertise of the end user of the explanations affects explainability; and

(e) Mismatch of expectations particularly affects trust.

5.2   Limitations and future work

There are a few limitations:

1. *Limited number of respondents (n = 31), diversity of expertise and age*:
   The original intent was to conduct a more in-depth cross-sectional
   study of the self-reported scores of the respondents according to
   their area of expertise and their beliefs relating to AI and data pri-
   vacy. This analysis could be used to support certain inferences
   made earlier that explainability is dependent on expertise of the
   end user. However, with 31 respondents, breaking down the survey
   results into smaller sub-groups would not reliably show statistical
   trends[1]. Hence, future work could be conducted on a much larger
   sample size to run more cross-sectional analysis.

2. *Analysis is limited to the context of the APP-350 corpus, classifiers used,
   and LIME:* Assessing XAI is highly context dependent, and therefore
   the foregoing analysis should also be seen in light of these partic-
   ular components. Future work to evaluate XAI within the context
   of data privacy could include using another data privacy dataset,
   other classifiers and other XAI techniques apart from LIME. In par-
   ticular, LIME may interact differently with current state-of-the-art
   classifiers which are much more complex and rely on deep learning
   given that NLP has developed quite substantially since 2016 when
   LIME was first introduced.

---

[1]For example, Wilcoxon Rank Sum Test is recommended to be run on populations
> 20.

3. *Inherent drawbacks of LIME and human evaluation of XAI:* As LIME is a post-hoc local explainability technique, respondents' understanding of the model was limited to the example visualisations that were presented to them. The sentences shown to respondents were intended to show the limitations of the classifier. This paints a more confusing and inconsistent picture of the classifier's logic, as compared to examples that consistently show "cookie" was the most important feature. In fact, this was controlled for in another study where the authors only chose to show correctly classified sentences in order to reduce the information load on the respondents [11]. Hence, if more consistent examples were chosen instead, respondents might have responded differently. If post-hoc techniques are used, future work could include an interactive dashboard where respondents can choose to generate their own sentences for classification, which allows them to test their understanding of the classifier's logic is consistent.

   Further, human evaluations of post-hoc explanations have been criticised for being inherently flawed because of confirmation bias. Post-hoc explanations may have little to no similarities to how the underlying classifier made the prediction as they are an oversimplification of the classifier's logic. Hence, human evaluation may have limited meaning [27]. Instead, the authors propose four objective metrics that quantify the explanation itself and its appropriateness given the XAI goal. Future work could involve less reliance on human evaluation to incorporate more objective metrics, as well as

choosing other types of XAI techniques such as global self explaining techniques. Future work could also investigate assigning objective explainability ratings to AI models [34], similar to car safety ratings.

4. *Investigation of the differences in interaction between explainability metrics and other areas of algorithmic legal decision making:* The explainability metrics used in Section 4.1 were presumed to be important values for a legal context because of their link to the rule of law. Nevertheless, there are other values that could be relevant such as cost because of access to justice and predictability because the legal system should provide consistent rules for its subjects to follow. Further investigation could also be conducted on how these values change and affect explainability in different areas (i.e. tort vs data privacy) and fora (i.e. PDPC vs the Court of Appeal) of the law because of different legal standards [14] and the expectations & pre-conceptions of users of the legal system [35].

In particular, as the intuition behind increasing explainability is to increase trust of users, it is worth further investigating this relationship between explainability and trust. There has been evidence that shows this relation is not as general as previously thought [18]. In fact, if explanations reveal problems about the system, trust could decrease rather than increase, which seems to be the case in this capstone where respondents were intentionally given examples that demonstrated the limitations of the classifiers.

5.3 Final comments

Ultimately, the application of XAI and the law is a multi-disciplinary field that requires both empirical and normative investigation from the different disciplines of data science, law, psychology, philosophy, social sciences and more. This capstone only scratches the surface of the broader issues of XAI and the law, including defining legal standards for explainability and translating them into technical XAI requirements. To this end, the AI Act proposed by the European Commission differentiates AI systems into those which are "high risk" and "medium risk", with "high risk" systems requiring a greater level of explainability that is user-empowering and compliance-oriented [29], though the details and implementation are still under discussion.

Thus, even the law is not exempt from the winds of change despite its long tradition and established processes. With suggestions that GPT-4 has the markings of artificial general intelligence [5], and Google's $400 million investment in AI firm Anthropic in response to Microsoft's $10 billion investment into OpenAI [1], algorithmic decision making as a whole is not only going to be more pervasive, but also likely more opaque. Law, as an institution centered on explanation and justification through argumentation, should also be equally concerned with the advancement of XAI.

Bibliography

[1]   Davey Alba and Dina Bass. *Google Invests Almost 400 million in Chat-GPT Rival Anthropic*. Last Accessed: 2023-03-28. 2023. URL: https://www.bloomberg.com/news/articles/2023-02-03/google-invests-almost-400-million-in-ai-startup-anthropic#xj4y7vzkg.

[2]   Hussam Alkaissi and Samy I McFarlane. "Artificial hallucinations in chatgpt: Implications in scientific writing". In: *Cureus* 15.2 (2023).

[3]   Kevin D. Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017. DOI: 10.1017/9781316761380.

[4]   Patricia Bassey. *Logistic Regression vs Support Vector Machines (SVM)*. Last Accesssed: 2023-03-21. 2019. URL: https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16.

[5]   Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712 [cs.CL].

[6]   Simon Chesterman. "We, The Robots: Regulating Artificial Intelligence and the Limits of the Law". In: Cambridge University Press, 2021. Chap. 6. Transparency.

[7]   Simon Chesterman. "We, The Robots: Regulating Artificial Intelligence and the Limits of the Law". In: Cambridge University Press, 2021. Chap. 3. Opacity.

[8]   Marina Danilevsky et al. "A Survey of the State of Explainable AI for Natural Language Processing". In: *CoRR* abs/2010.00711 (2020). arXiv: 2010.00711. URL: https://arxiv.org/abs/2010.00711.

[9]   Finale Doshi-Velez and Been Kim. "Towards a Rigorous Science of Interpretable Machine Learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[10]   Julia El Zini and Mariette Awad. "On the Explainability of Natural Language Processing Deep Models". In: *ACM Computing Surveys* 55.103 (2022), pp. 1–31. DOI: https://doi.org/10.1145/3529755.

[11]   Łukasz Górski and Shashishekar Ramakrishna. "Explainable Artificial Intelligence, Lawyer's Perspective". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. São Paulo, Brazil: Association for Computing Machinery, 2021, 60–68. ISBN: 9781450385268. DOI: 10.1145/3462757.3466145. URL: https://doi.org/10.1145/3462757.3466145.

[12]   Stanley Greenstein. "Preserving the rule of law in the era of artificial intelligence (AI)". In: *Artificial Intelligence and Law* 30.3 (2022), pp. 291–323.

[13]   Oskar J. Gstrein and Anne Beaulieu. "How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches". In: *Philos Technol* 35.1 (2022), p. 3. DOI: https://doi.org/10.1007%2Fs13347-022-00497-4.

[14] Philipp Hacker and Jan-Hendrik Passoth. "Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond". In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer. 2022, pp. 343–373.

[15] Bi-Min Hsu. "Comparison of supervised classification models on textual data". In: *Mathematics* 8.5 (2020), p. 851.

[16] IBM. *What is natural language processing?* Last Accessed: 2023-02-11. 2022. URL: https://www.ibm.com/sg-en/topics/natural-language-processing.

[17] Javatpoint. *Logistic Regression in Machine Learning*. Last Accesssed: 2023-03-21. n.d. URL: https://www.javatpoint.com/logistic-regression-in-machine-learning.

[18] Lena Kästner et al. "On the relation of trust and explainability: Why to engineer for trustworthiness". In: *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE. 2021, pp. 169–175.

[19] Daniel Martin Katz et al. "GPT-4 Passes the Bar Exam". In: *Available at SSRN 4389233* (2023).

[20] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy (Basel)* 23.1 (2020), p. 18. DOI: https://dx.doi.org/10.3390/e23010018.

[21] Zhiyuan Liu et al. "Word representation". In: *Representation Learning for Natural Language Processing* (2020), pp. 13–41.

[22] Alessandro Mantelero. "The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics". In: *Computer Law & Security Review* 30.6 (2014), pp. 643–660.

[23] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. Last Accessed: 2023-02-11. 2022. URL: https://openai.com/blog/chatgpt/.

[24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[25] Personal Data Protection Commission. *Model Artificial Intelligence Governance Framework Second Edition*. Tech. rep. Last Accessed: 2023-2-20. Personal Data Protection Commission, Jan. 2020. URL: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. DOI: 10.48550/ARXIV.1602.04938. URL: https://arxiv.org/abs/1602.04938.

[27] Avi Rosenfeld. "Better Metrics for Evaluating Explainable Artificial Intelligence". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. Virtual Event, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2021, 45–50. ISBN: 9781450383073.

[28] SciPy. *scipy.stats.wilcoxon*. Last Accessed: 2023-3-1. 2023. URL: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html`.

[29] Francesco Sovrano et al. "Metrics, explainability and the European AI act proposal". In: *J* 5.1 (2022), pp. 126–138.

[30] Daniel Vale, Ali El-Sharif, and Muhammed Ali. "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law". In: *AI and Ethics* (2022), pp. 1–12.

[31] Giulia Vilone and Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence". In: *Information Fusion* 76 (2021), pp. 89–106.

[32] Ekaterina Vylomova et al. *Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning*. 2015. DOI: `10.48550/ARXIV.1509.01692`. URL: `https://arxiv.org/abs/1509.01692`.

[33] Isabel Wagner. "Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021". In: *arXiv preprint arXiv:2201.08739* (2022).

[34] Bernhard Waltl and Roland Vogl. "Explainable artificial intelligence the new frontier in legal informatics". In: *Jusletter IT* 4 (2018), pp. 1–10.

[35] Gizem Yalcin et al. "Perceptions of justice by algorithms". In: *Artificial Intelligence and Law* (2022), pp. 1–24.

[36]   Carlos Zednik. "Solving the black box problem: A normative frame-
        work for explainable artificial intelligence". In: *Philosophy & technol-
        ogy* 34.2 (2021), pp. 265–288.

[37]   Sebastian Zimmeck et al. "MAPS: Scaling Privacy Compliance Anal-
        ysis to a Million Apps". In: *Proceedings on Privacy Enhancing Tech-
        nologies* 2019.3 (2019), pp. 66–86.