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

DECLARATION & CONSENT

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 (Yale-NUS HR 039).

ACCESS LEVEL

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

   ο Unrestricted access
   Make the Thesis immediately available for worldwide access.

   ο Access restricted to Yale-NUS College for a limited period
   Make the Thesis immediately available for Yale-NUS College access only from _____________ (mm/yyyy) to _______________ (mm/yyyy), up to a maximum of 2 years for the following reason(s): (please specify; attach a separate sheet if necessary):
   ____.

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

   ο Other restrictions: (please specify if any part of your thesis should be restricted)
   ____
   ____

____
Name & Residential College of Student

____________________________________ ___________________________
Signature of Student Date

____________________________________ ___________________________
Name & Signature of Supervisor Date

Acknowledgements

他对我说: "我的恩典够你用的, 因为我的能力是在人的软弱上显得完全." 所以, 我更喜欢夸耀自己的软弱, 好使基督的能力覆庇我. - 哥林多后书12:9

Thanks be to God, who has provided me with countless opportunities to learn and grow academically and non-academically, and who has granted me the strength to overcome self-doubt to find joy in His love. His abundant grace has made this capstone (and more) possible.

I could not be more blessed to have my parents who have tirelessly and selflessly cared for me as my physical family, and to Law CF & 恩岭, who have extended their unconditional love to provide me with a spiritual home.

My heartfelt appreciation goes to my supervisors, Prof. Chesterman & Prof. Choi, who have willingly taken on this strange mixture of MCS and law and offered invaluable feedback and encouragement, and Prof. Danvy & Prof. Francesca, who have welcomed me into MCS and made me feel at ease despite my imposter syndrome.

I am privileged to have had journeyed with my friends Rayner, Nat, Ziying, and Glen into adulthood, and to have had the companionship of my DDP batchmates, who have made this five-year sojourn all the more transformative.

Finally, I thank my survey respondents, because without their participation, I would only have half a capstone.

When the race is complete, still my lips shall repeat: Yet not I, but through Christ in me!

YALE-NUS COLLEGE

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 are now much more advanced and can potentially automate some areas of substantive legal reasoning and writing, but has come at the cost of increased opacity of the models. Designing Explainable AI (XAI) is a method to combat this opacity, but it is unclear whether these explanations are effective across the different users and use cases of the models. The stakeholders in data privacy regulation can also benefit from the use of XAI. There is a gap in research of XAI assessments generally, 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 apps' data privacy practices. XAI techniques are applied on these models to produce visualisations that aim to explain these predictions to these stakeholders that may not be experts in data science. Finally, the effectiveness of these visualisations are assessed by surveying the general public along differing metrics of explainability.

"Any sufficiently advanced technology is indistinguishable from magic."

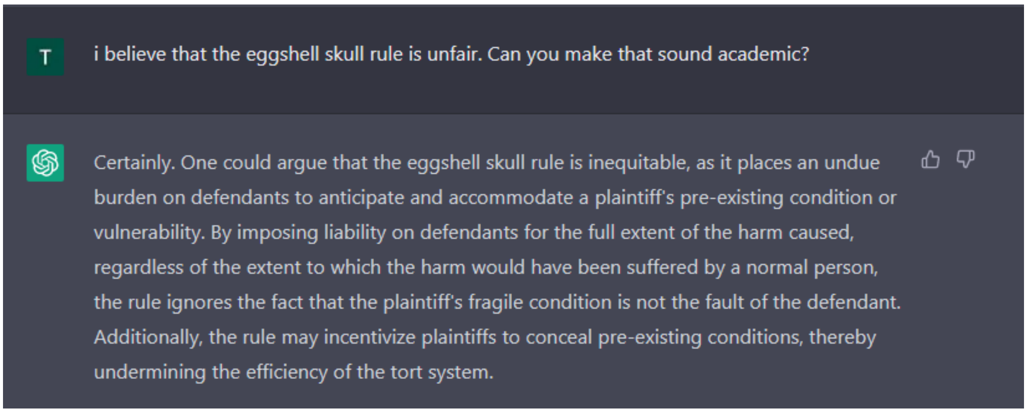
Arthur C. Clarke

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?" (Ashley, 2017). This question seems close to getting an answer in 2023 with the release of ChatGPT (OpenAI, 2022) to the general public. Consider the following response from ChatGPT:



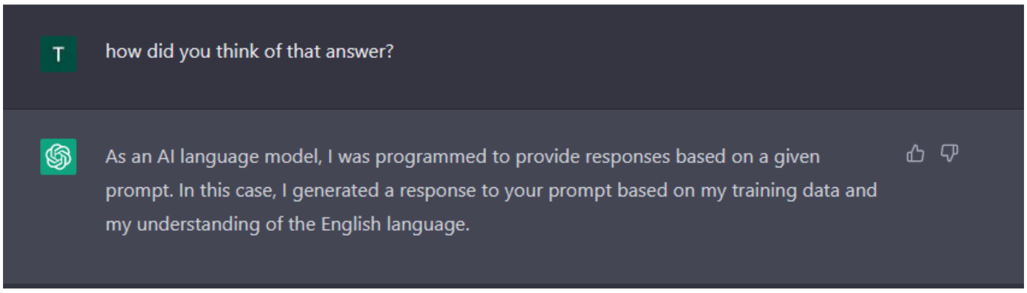
T i believe that the eggshell skull rule is unfair. Can you make that sound academic?

Certainly. One could argue that the eggshell skull rule is inequitable, as it places an undue burden on defendants to anticipate and accommodate a plaintiff's pre-existing condition or vulnerability. By imposing liability on defendants for the full extent of the harm caused, regardless of the extent to which the harm would have been suffered by a normal person, the rule ignores the fact that the plaintiff's fragile condition is not the fault of the defendant. Additionally, the rule may incentivize plaintiffs to conceal pre-existing conditions, thereby undermining the efficiency of the tort system.

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 (Katz et al., 2023).

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 (IBM, 2022). 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 (El Zini and Awad, 2022). This issue of explainability is exemplified by the last question I posed to ChatGPT about its view on the egg-shell skull rule:



how did you think of that answer?

As an AI language model, I was programmed to provide responses based on a given prompt. In this case, I generated a response to your prompt based on my training data and my understanding of the English language.

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

[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 (Alkaissi and McFarlane, 2023).

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

legal tech tools but does not understand how the analysis was produced could be considered lacking in diligence and competence.

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 (Linardatos, Papastefanopoulos, and Kotsiantis, 2020) 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 (Danilevsky et al., 2020). 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" (Danilevsky et al., 2020). 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 (Rosenfeld, 2021). 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 (Rosenfeld, 2021). 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 (Gstrein and Beaulieu, 2022). 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] (Wagner, 2022). 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 consumer giving consent that is informed and freely given (Mantelero, 2014). 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

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

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

the decision making process of the organisations so as to build understanding and trust with those stakeholders that use their products (Personal Data Protection Commission, 2020). 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.

Despite debate about the scope of the right to explanation and whether it is binding, (Chesterman, 2021a), 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 (Chesterman, 2021b) 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

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

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.

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") 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?

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

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[7]. 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.

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, 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 (Górski and Ramakrishna, 2021). 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 methodology of the selected NLP models, 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 closes the capstone with limitations and a general discussion of the findings in relation to the research questions.

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

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 5 data practices 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 (Zimmeck et al., 2019). 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

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

of Krippendorff's $\alpha = 0.78$[2], the agreement between the annotations exceeded previous similar research.

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.

2.1.1 Data pre-processing

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] 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.

[3] A segment is similar to a paragraph.

2.2 Text representations and classifiers

2.2.1 Literature review

In NLP, 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 (Liu et al., 2020). 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(Liu et al., 2020). 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 (Zednik, 2021). In generating text representations of the APP-350 corpus, Zimmeck et al. 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 models 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, Zimmeck et al. used support vector machine for classification (SVCs) to achieve reasonable performance. Individual classifiers 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, Zimmeck et al. 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].

[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.

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 for prediction in this capstone. Since sparse vector representations and classical models have performed well for the APP-350 corpus 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: $\text{King} - \text{Man} + \text{Woman} = \text{Queen}$ (Vylomova et al., 2015). GloVe is an unsupervised learning algorithm for obtaining vector representations for words.

[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.

The length-50 vectors used in this capstone are pre-trained using this algorithm on Wikipedia and news articles (Pennington, Socher, and Manning, 2014). 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 this capstone, 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 in the case of Tf-IDF. 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. I also use logistic regression 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 the capstone.

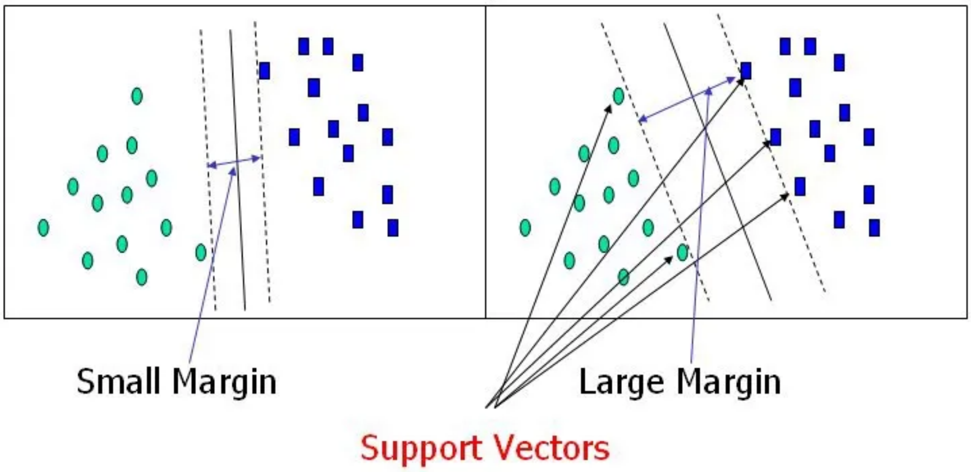


FIGURE 2.1: Finding the hyperplane in SVC (Bassey, 2019).

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.

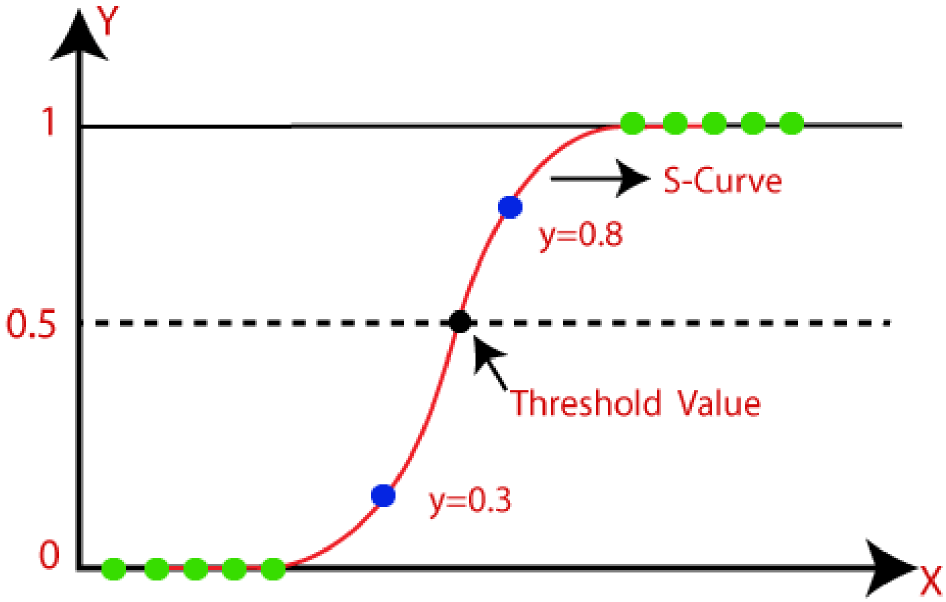


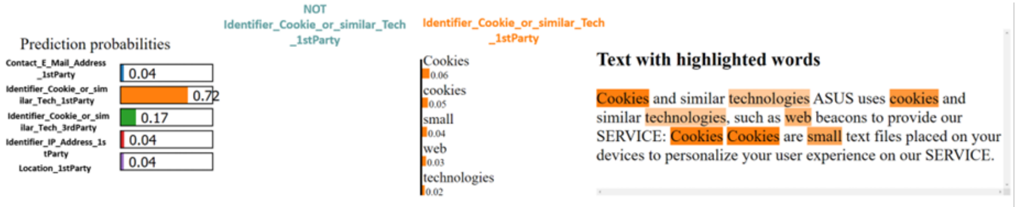
FIGURE 2.2: Sigmoid function in logistic regression (Javatpoint, n.d.).

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 Danilevsky et al., 2020.

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.


Prediction probabilities
Contact_E_Mail_Address_1stParty
0.04
Identifier_Cookie_or_similar_Tech_1stParty
0.72
Identifier_Cookie_or_similar_Tech_3rdParty
0.17
Identifier_IP_Address_1stParty
0.04
Location_1stParty
0.04
NOT Identifier_Cookie_or_similar_Tech_1stParty
Identifier_Cookie_or_similar_Tech_1stParty
Cookies
0.06
cookies
0.05
small
0.04
web
0.03
technologies
0.02

Text with highlighted words

Cookies and similar technologies ASUS uses cookies and similar technologies, such as web beacons to provide our SERVICE: Cookies Cookies are small text files placed on your devices to personalize your user experience on our SERVICE.

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 (Ribeiro, Singh, and Guestrin, 2016). 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 (Ribeiro, Singh, and Guestrin, 2016).

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. Standards of explainability are highly dependent on the purpose of generating the explanations. A few purposes have been suggested (Doshi-Velez and Kim, 2017):

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 (Vale, El-Sharif, and Ali, 2022) might not apply *mutatis mutandis* to the GDPR. In any case, this capstone makes no claim as to the applicable standard for explainability for data privacy, and merely serves as empirical insight as to the perceptions of explainability by potential data subjects.

In terms of methods of assessing XAI, human evaluation has been frequently utilised. (Danilevsky et al., 2020). 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 (Doshi-Velez and Kim, 2017).

Human evaluation has been conducted within a legal context (Górski and Ramakrishna, 2021). 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 (Rosenfeld, 2021). 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 judged according to such values (Greenstein, 2022). 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 and tweak the methodology of Gorski and Ramakrishna. 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. **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. 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, is a risk as well as how far they are concerned about their data privacy. This is used to understand

[6] The full survey can be found here, and a summary of the responses can be found here.

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. **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) to measure the variability of different values related to explainability across different purposes of explanation. All three contexts relate to using the classifier to make a decision with legal consequences, and are meant to vary the purposes of the explanation according to the three purposes suggested by Doshi-Velez & Kim above. After reading each context, the respondents are 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, and are used to measure how these values that relate to explainability can vary according to the purposes of the explanation.

3. **Rating of the LIME visualisations of four classifications by the same classifier (Logistic regression + Tf-IDF).** 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. 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 metric to measure explainability, aside from self-reported scores.

4. **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 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). The respondents' choices are then totalled up and compared against traditional performance metrics (P/R/F1) of the classifiers, to investigate whether there is a relation between interpretability and model performance.

5. **Repeating the questions for the same three contexts.** Respondents are asked to view the same three contexts and answer the same questions in Part 2. This evaluates whether viewing the explanations significantly changed the respondents' beliefs according to the four metrics, and whether there are intra-relations and trends of the individual metrics and contexts.

The survey was hosted on Google Forms, and respondents were 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.

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 (Table 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 3.2) and as seen by similar interquartile range (IQR) (Figure 3.2c). 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 3.1) that cause a slightly right skew distribution (Figure 3.2a). The distributions are also unimodal, with the mode around 20.

practice	counts	mean sentence length	median sentence length	percentage of total
Identifier_Cookie_or_similar_Tech_1stParty	2107	25.4	22.0	11.2%
Contact_E_Mail_Address_1stParty	2106	28.7	25.0	11.2%
Location_1stParty	1514	29.2	24.0	8.1%
Identifier_Cookie_or_similar_Tech_3rdParty	1250	27.3	24.0	6.6%
Identifier_IP_Address_1stParty	1005	30.9	27.0	5.3%
Contact_Phone_Number_1stParty	970	29.1	25.0	5.2%
Identifier_Device_ID_1stParty	697	32.4	28.0	3.7%
Contact_Postal_Address_1stParty	597	28.9	26.0	3.2%
SSO	504	32.6	28.0	2.7%
Demographic_Age_1stParty	428	33.1	26.0	2.3%

TABLE 3.1: Summary statistics per practice of top 10 occurring practices.

	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 3.2: Summary statistics of top 10 occurring practices.

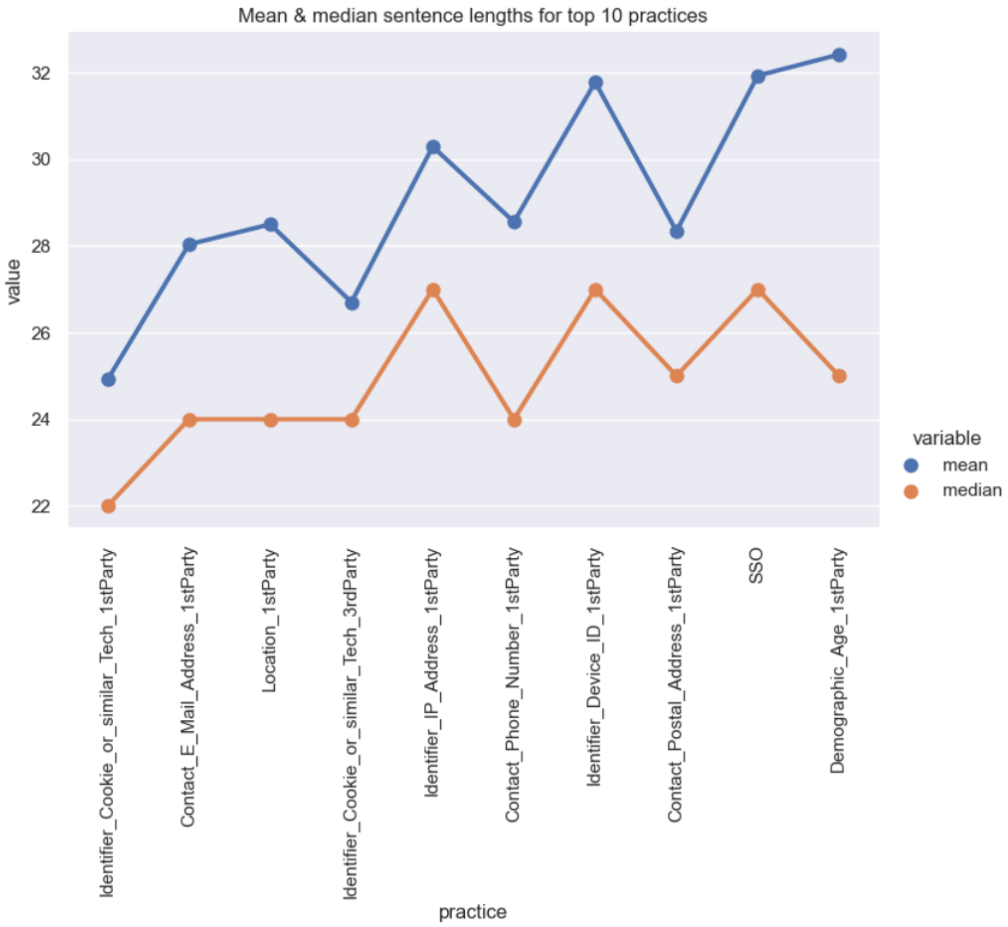
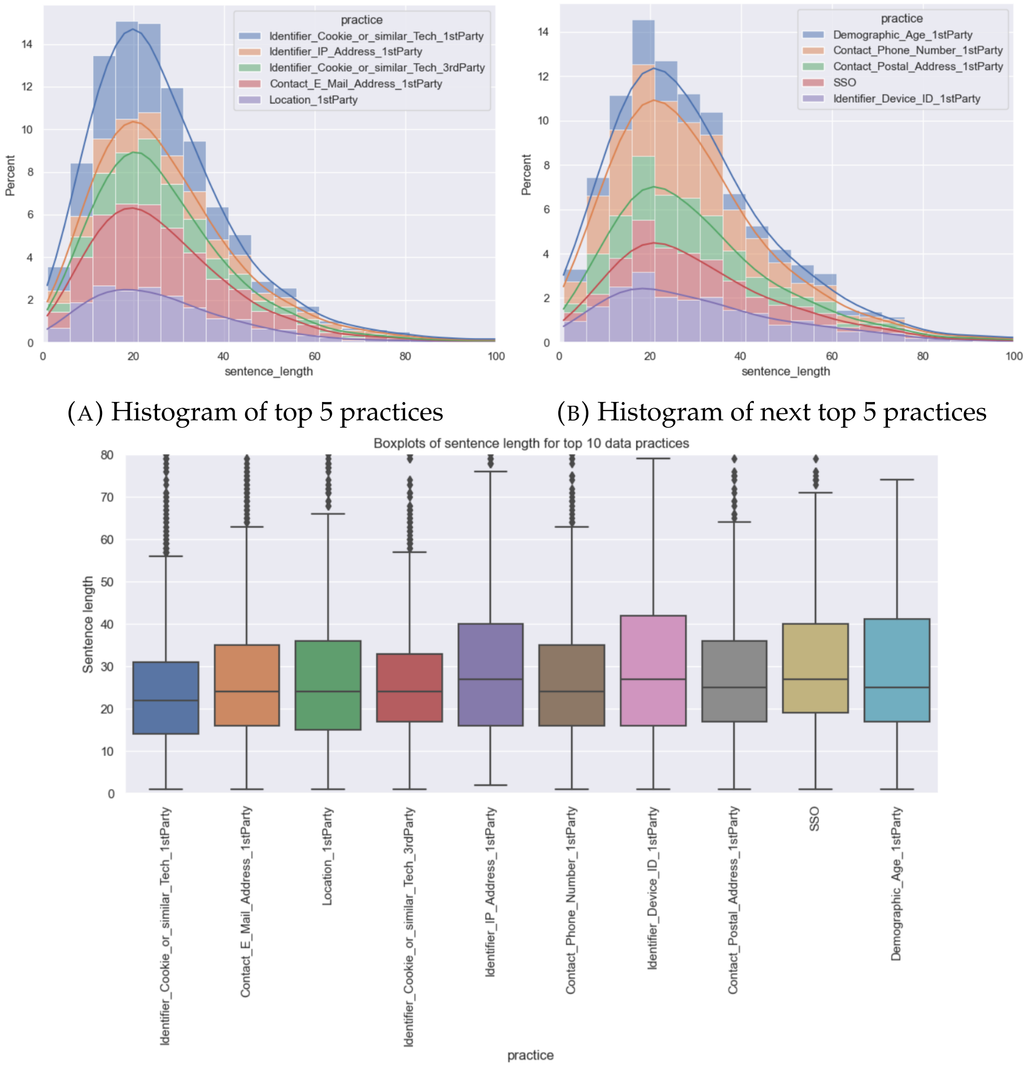


FIGURE 3.1: Mean and median sentence length

With regard to top 10 frequently occurring 4-grams in the top 5 practices (Figure 3.3), both 4-grams of `Cookie 1st Party` and `Cookie 3rd`



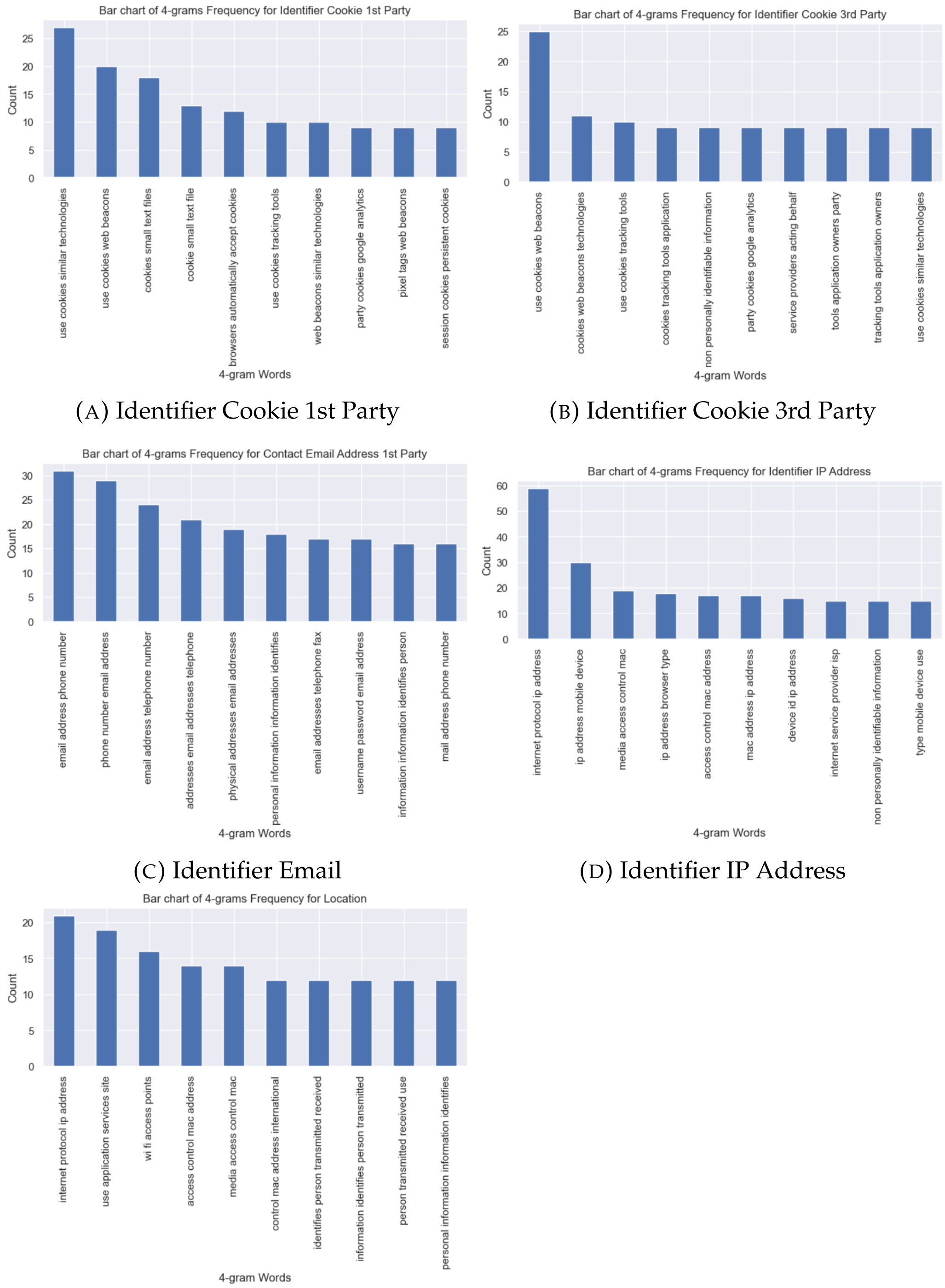
(A) Histogram of top 5 practices

(B) Histogram of next top 5 practices

(C) Boxplots of sentence length (Practices shown in order of descending frequency)

FIGURE 3.2: Summary plots for top 10 frequently occurring data practices (note: not all outliers shown to focus on visualising IQR)

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



(A) Identifier Cookie 1st Party

(B) Identifier Cookie 3rd Party

(C) Identifier Email

(D) Identifier IP Address

(E) Location 1st Party

FIGURE 3.3: Top 10 most frequently occurring 4-grams in the top most frequently occurring 5 data practices at the sentence level

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 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 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 3.4). This corresponds with the findings by Zimmeck et al. 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. 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.

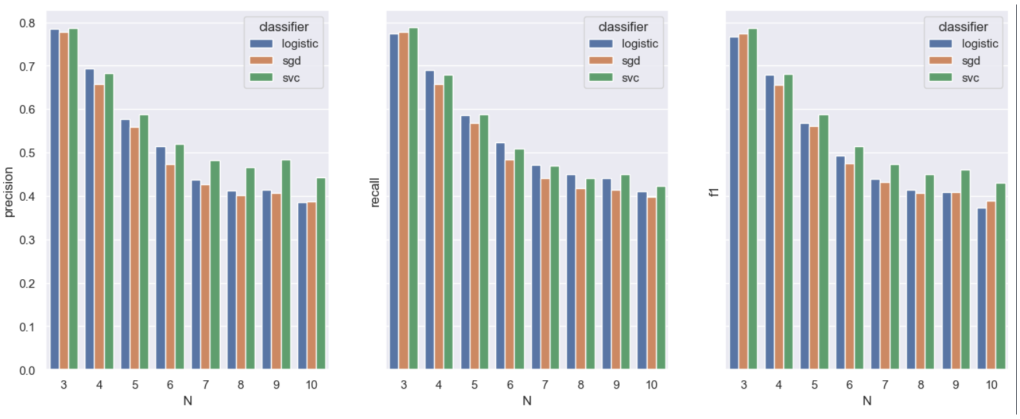


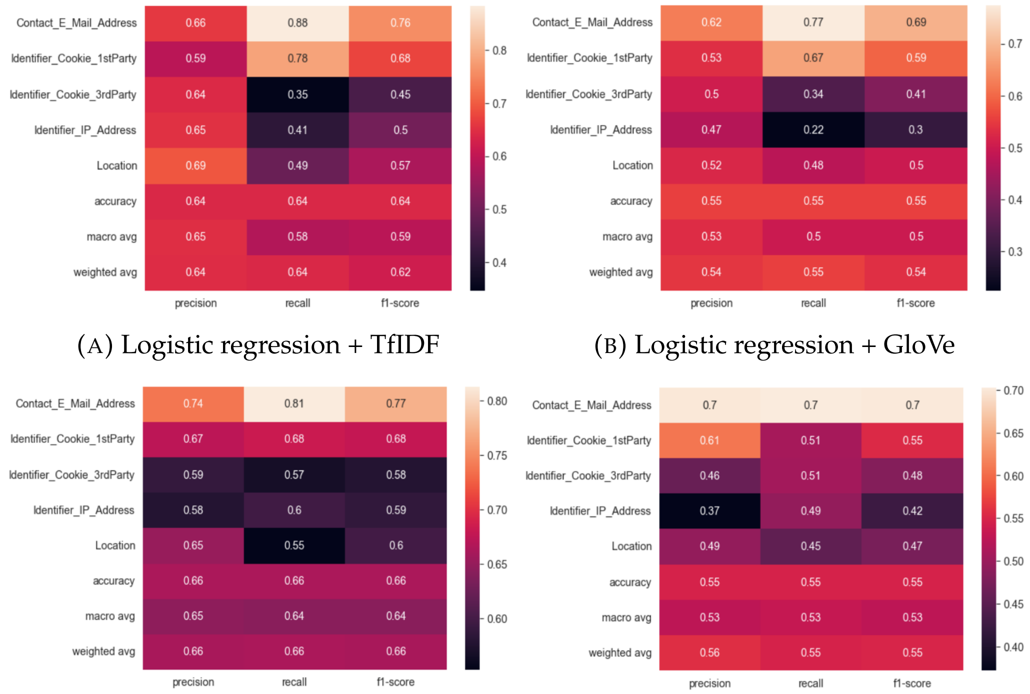
FIGURE 3.4: Weighted average P/R/F1 scores for top N ($3 \leq N \leq 10$) practices

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 3.5), similarly to what Zimmeck et al. 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 3.6), 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 (Hsu, 2020), where LinearSVC followed by logistic regression were the highest performing models.



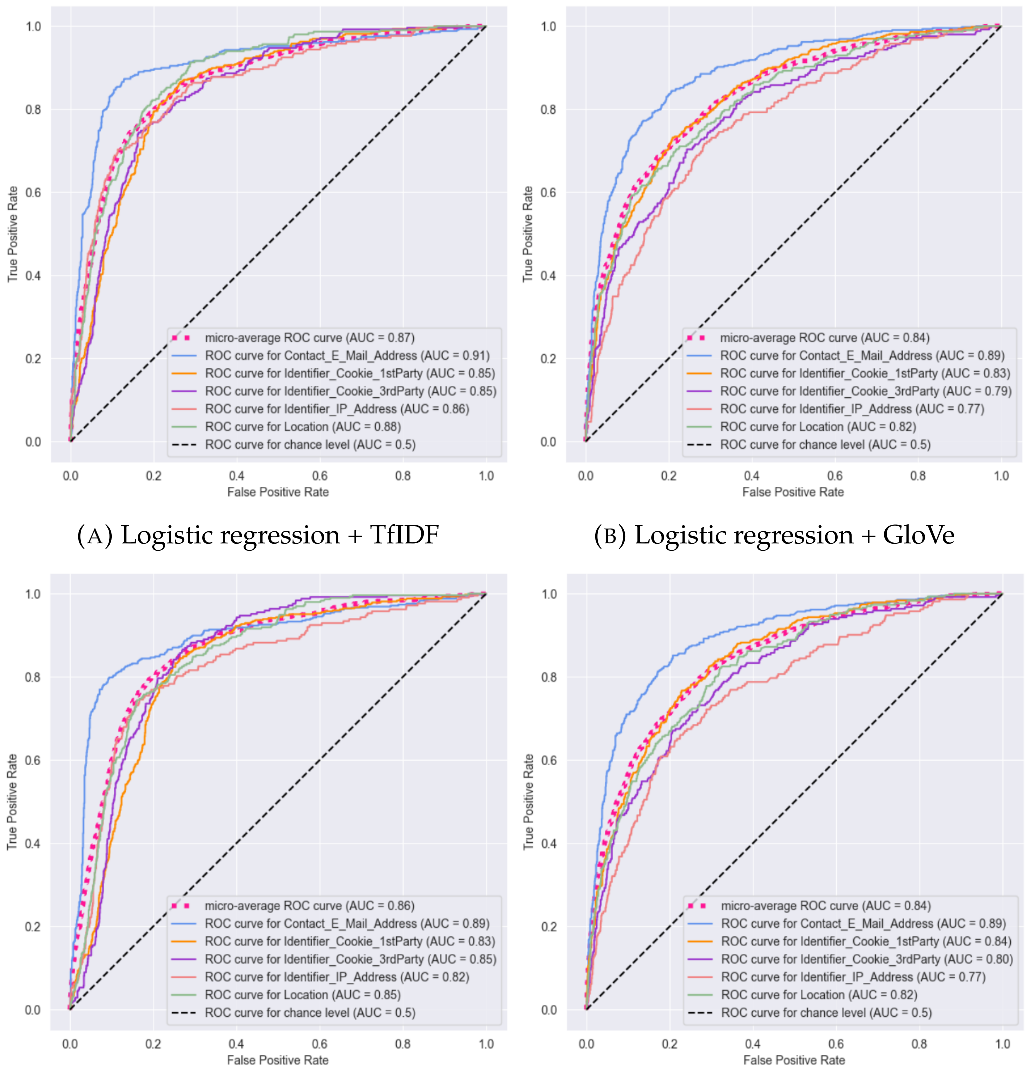
(A) Logistic regression + TfIDF

(B) Logistic regression + GloVe

(C) SVC + TfIDF

(D) SVC + GloVe

FIGURE 3.5: 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 3.6: Multiclass Receiver Operating Characteristic (ROC) Curve of the 5 data practices for each classifier

4 Discussion of survey results

4.1 Summary of results

In total, 31 responses were collected. The majority of respondents majored in law (58%) vs non-law (42%) respondents. Figure 4.1 and Figure 4.2 show the demographic data and responses relating to the respondents' beliefs regarding AI and data privacy. I originally intended to conduct analysis of the results across different demographics such as those with more expertise in computer science, or those who think that decisions by AI are not as useful. However, 31 respondents are insufficient for further meaningful demographic breakdown.

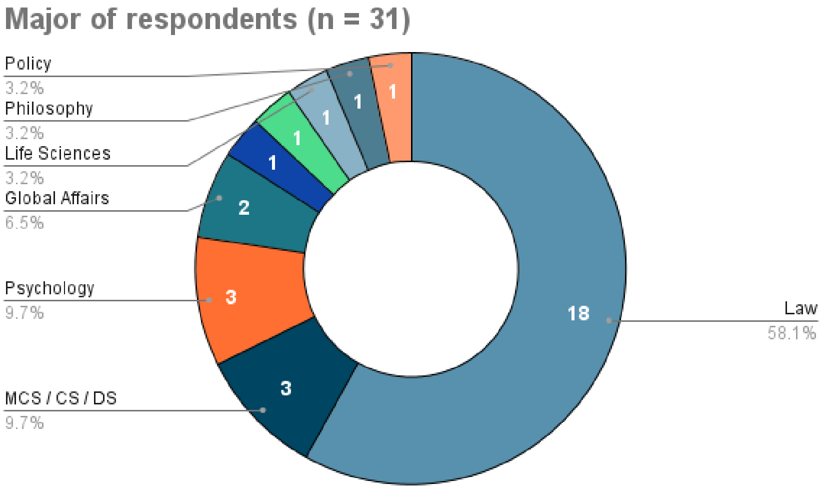


FIGURE 4.1: Breakdown of respondents' expertise by major

4.1.1 Part 1: Beliefs relating to AI & data privacy

Except for the questions relating to subject matter expertise (data privacy and AI), the level of agreement of law vs non-law respondents 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.

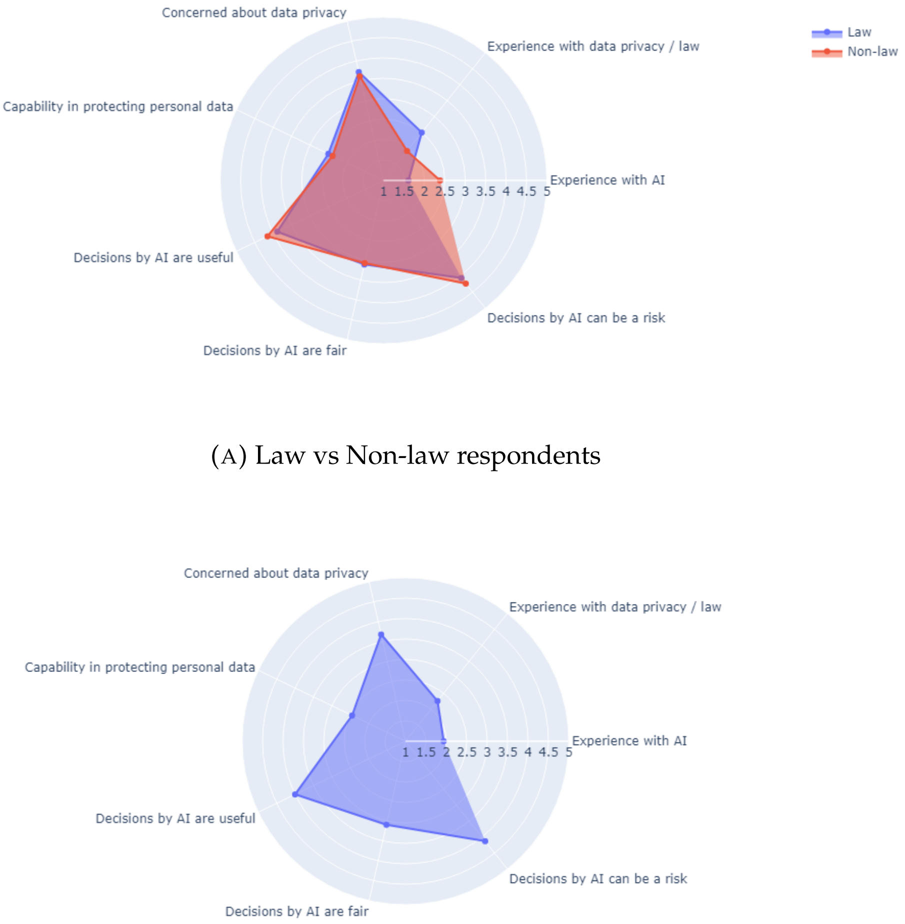
4.1.2 Part 2 & Part 6: Comparison of self-reported scores of explainability across the three contexts

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 explainations.

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

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 (SciPy, 2023). 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 $d < 0$,



(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. $n = 31$)

then $i < f$ and the scores increased after viewing. Conversely, if $d > 0$, then $i > f$ and the scores decreased after viewing.

The p-values of the tests are reported in Table 4.1 , 4.2 and **??**. For Table 4.2a and 4.2b, I took the mean of self-reported scores across contexts, and across each metric, respectively. "Increase" and "decrease" refer

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) Mean scores by metrics

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) Mean scores by contexts

Increase	Decrease
0.844	0.156

(C) Mean scores across contexts and metrics

TABLE 4.2: p-values comparing mean scores

to the p-values of the 1-sided test that checks whether the scores significantly increased or decreased after viewing the explanations. For Table **??**, the self-reported scores were averaged across both context and metric. Here are the instances when $p < 0.1$ (highlighted in red in the tables), and therefore $H0$ can be rejected in favour of $H1$ such that there was a statistically significant increase / decrease of scores after viewing the explanations:

1. Effectiveness and trust significantly decreased for context 2 (PDPC), and fairness significantly decreased for context 3 (consumer). Only

effectiveness significantly increased for context 1 (app developer) (Table 4.1). While trust technically did not have a statistically significant decrease for context 1, the p-value is quite close at 0.113. Hence, there seems to be an inverse relationship between effectiveness and trust for context 1, but a positive relationship for context 2.

2. There is a statistically significant decrease in explainability metrics for context 1 and 2 when comparing the mean of the scores across the metrics (Table 4.2a) and for trust across the three contexts (Table 4.2b).

3. The overall self-reported explainability when taking the mean of the scores across both contexts and metrics did not significantly increase / decrease (Table **??**).

Remember that the explanations given to respondents were deliberately chosen to demonstrate the limitations of the model. Hence, it can be inferred that respondents were more cognisant of such limitations when they answered the same questions again in Part 6. Here are some inferences that can be drawn from these observations:

1. The inverse relationship between effectiveness and trust for context 1 (app developer) could be explained by the fact that respondents believed that software engineers may be more concerned with a standardised, consistent way to identify data practices, as someone who is not legally trained but is still subjected to data privacy obligations. Hence, self-reported effectiveness increased when they realised that the model would still do a better job at identifying data

practices consistently compared to themselves. However, trust decreased because respondents also knew that the model could make wrong predictions. In this context, respondents were willing to bear some risk of getting predictions wrong in return for automated detection of data practices.

2. For the positive relationship between effectiveness and trust for context 2 (PDPC), respondents were more concerned about retaining discretion in their decision making process. Since they knew that the model could make wrong predictions, combined with the fact that PDPC is the regulatory body that enforces data privacy, the costs of making a wrong decision is extremely high. Hence, even though the model is more "effective" since it is automated, it is not "effective" in terms of making a legally defensible decision given that it is fallible. Trust decreased because while the respondents believed that this model could make their jobs in identifying data breaches less repetitive (in that they would not have to read every data privacy policy manually), there is now a mismatch of expectations because respondents still have to ensure that the predictions given by the model are legally correct.

3. For the consumer, as a person who is not trained in law or data science, they are subjected to how the app developer designed the app such that the app developer failed to disclose that the app was tracking cookies. However, as consumers are not legally trained, they cannot correct this breach of their privacy rights except by relying on the fallible predictions of the model. There is a chance that the model makes the wrong predictions, and the respondent fails to

make a case to the PDPC. Therefore, it is not "fair" to the consumer that though their privacy was violated, they cannot do anything about it because of their reliance on the fallible model.

4. Overall, a different weighting of these four values by the respondents can be inferred from the significance tests. The respondents were more concerned about having automated identification of data practices as a software engineer, over the risk of the model making a wrong prediction. In this case, effectiveness (in the sense of efficiency of identifying data practices) was prioritised above trust. This is contrasted with the PDPC context in which respondents were highly concerned of the risks of making a wrong decision more so than automating reading data privacy policies, and so fallible models are seen as impacting effectiveness and trust towards making sound legal decisions. The consumers are mainly concerned with their own interests and whether their rights would be vindicated, and so fallible models would greatly hinder them from pursuing this self-interest which is unfair for them.

5. As to the comparison of aggregated scores in Table 4.2a, the fallibility of the model weighed heavily in the respondents' minds to reduce the explainability metrics in context 2 (PDPC) and context 3 (consumer). Perhaps the respondents thought that as app developers who were trained in programming, they would possess the requisite expertise to understand and work around the model's limitations compared to context 2 (PDPC) and context 3 (consumer) who may not have such technical knowledge and so are limited by the explanations that were presented to them.

6. In the case of trust being the only metric that significantly decreased across the contexts, this perhaps shows the interrelation between of model transparency and trust. Intuitively, if respondents do not understand how a process works, they are less likely to trust the results of that process. However, since the respondents were made aware of the limitations of the model, they still have reduced trust not because they did not understand the model, but that their expectations of the model in producing objective, 100% reliable predictions have been affected. Hence, there can be two plausible reasons for the loss in trust: the explanations were not effective in explaining the model, or that their expectations in the model have decreased as a result of being made aware of the limitations.

4.1.3 Part 3: Testing whether viewing more visualisations increase explainability

Analysis of the reported scores of "interpret" and "understand"

There is a decreasing trend of both understanding and interpret after viewing each explanation, though understanding has a more negative gradient than interpret (Figure 4.3). It seems that respondents found the model less generally explainable after viewing more explanations, both in terms of the method of visualising the explanation (which is the "interpret" question), as well as understanding the inner logic of the model (the "understand" question).

One interpretation of this decreasing trend suggests both the model and LIME were not very explainable. However, another interpretation could be that the respondents being more 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 in Chapter 2, the visualisations for this part were specifically chosen to demonstrate the limits of the model 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 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. This decrease in explainability could be likened to a variation of the Dunning-Kruger effect (Dunning, 2011), whereby the respondents who initially had no experience with the model believed that they understood the model well because they were unaware of the complexity of the model, but when they were given more information, they had more awareness of the complexity and therefore reported a drop in explainability of the model as a whole. Therefore, LIME and the model could be said to be locally explainable, but globally not explainable. If this interpretation of the results is adopted, it would not be surprising since LIME is designed as a post-hoc, local XAI technique.

Another interpretation is that the underlying explainability method also influences the respondents' perception of the explainability 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, explainability 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.

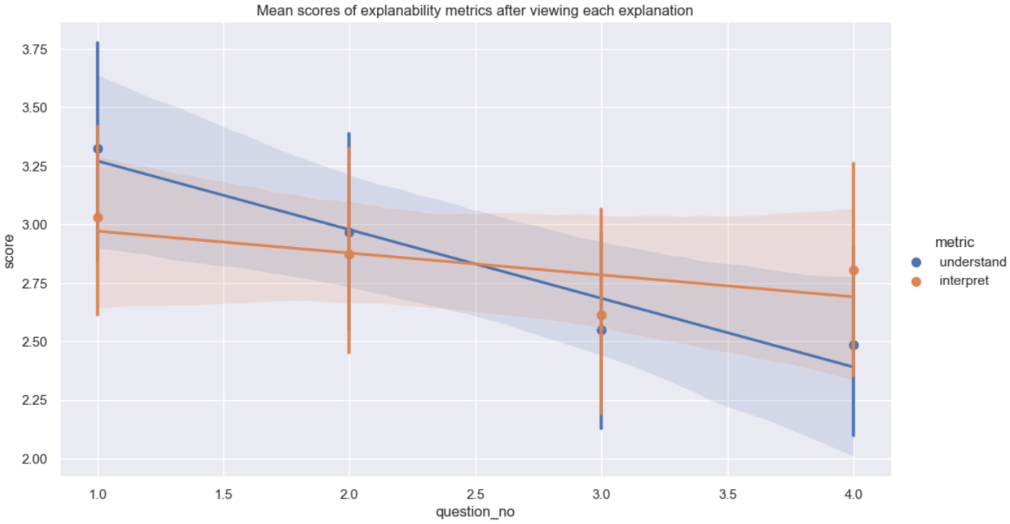
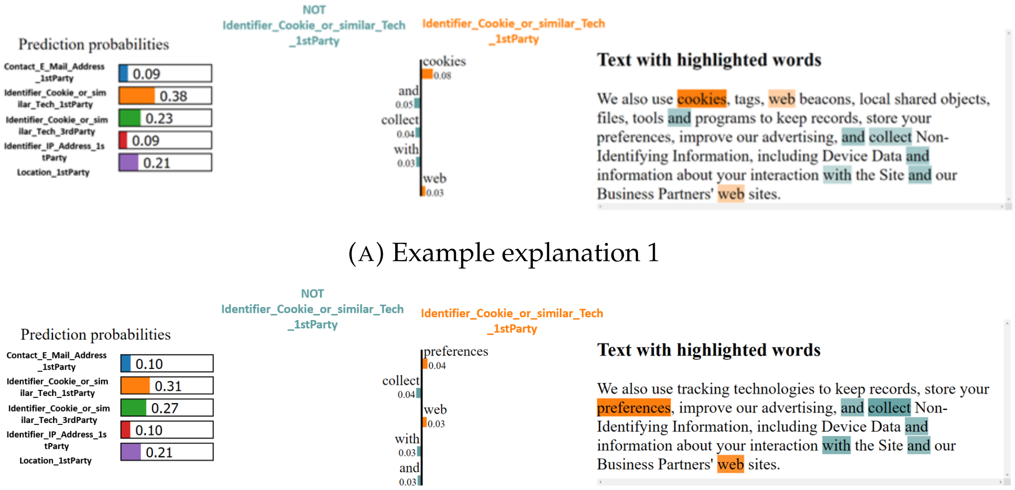


FIGURE 4.3: Trend of the mean of self-reported understanding and interpretability after viewing each explanation

Analysis of predicting counterfactuals

It should be noted that the model itself gave quite ambivalent predicted probabilities for each data practice. For example, for the second counterfactual (Figure 4.5), the model predicted `Contact_Email_Address` at 25% while `Identifier_Cookie` was predicted at 23%. Hence, the respondents' votes of predicting whether these counterfactuals would be classified as `Identifier_Cookie` should also be roughly ambivalent if the respondents actually had a good sense of how the model would predict. Only the results of the first question had some similarity to the predicted probabilities of the model, while the results of the third question differed the most from the model (Figure 4.7). Nevertheless, given the lack of any consistent trend across the questions, it is difficult to draw any definitive interpretations from the respondents' predictions of counterfactuals.


Prediction probabilities
Contact_E_Mail_Address_1stParty 0.09
Identifier_Cookie_or_similar_Tech_1stParty 0.38
Identifier_Cookie_or_similar_Tech_3rdParty 0.23
Identifier_IP_Address_1stParty 0.09
Location_1stParty 0.21
NOT Identifier_Cookie_or_similar_Tech_1stParty
Identifier_Cookie_or_similar_Tech_1stParty
cookies 0.08
and 0.05
collect 0.04
with 0.03
web 0.03
Text with highlighted words

We also use cookies, tags, web beacons, local shared objects, files, tools and programs to keep records, store your preferences, improve our advertising, and collect Non-Identifying Information, including Device Data and information about your interaction with the Site and our Business Partners' web sites.

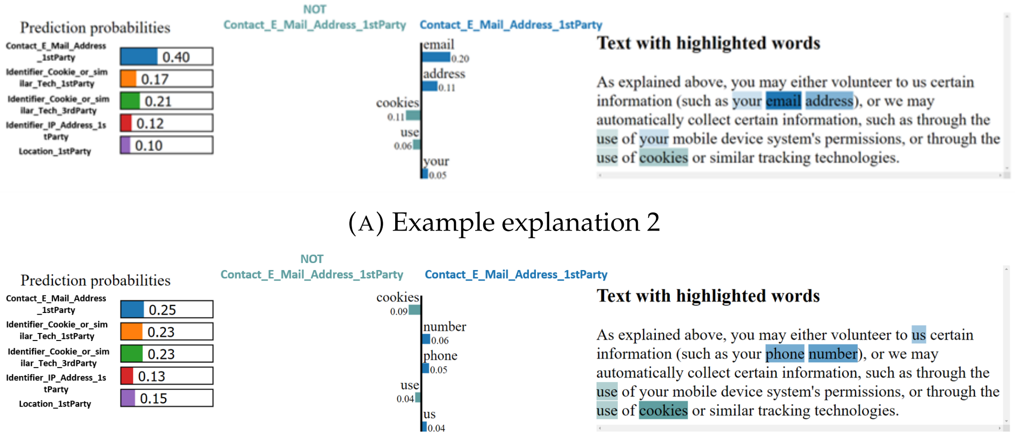
(A) Example explanation 1

Prediction probabilities
Contact_E_Mail_Address_1stParty 0.10
Identifier_Cookie_or_similar_Tech_1stParty 0.31
Identifier_Cookie_or_similar_Tech_3rdParty 0.27
Identifier_IP_Address_1stParty 0.10
Location_1stParty 0.21
NOT Identifier_Cookie_or_similar_Tech_1stParty
Identifier_Cookie_or_similar_Tech_1stParty
preferences 0.04
collect 0.04
web 0.03
with 0.03
and 0.03
Text with highlighted words

We also use tracking technologies to keep records, store your preferences, improve our advertising, and collect Non-Identifying Information, including Device Data and information about your interaction with the Site and our Business Partners' web sites.

(B) Counterfactual explanation 1

FIGURE 4.4: Example and counterfactual explanation 1


Prediction probabilities
Contact_E_Mail_Address_1stParty 0.40
Identifier_Cookie_or_similar_Tech_1stParty 0.17
Identifier_Cookie_or_similar_Tech_3rdParty 0.21
Identifier_IP_Address_1stParty 0.12
Location_1stParty 0.10
NOT Contact_E_Mail_Address_1stParty
Contact_E_Mail_Address_1stParty
email 0.20
address 0.11
cookies 0.11
use 0.06
your 0.05
Text with highlighted words

As explained above, you may either volunteer to us certain information (such as your email address), or we may automatically collect certain information, such as through the use of your mobile device system's permissions, or through the use of cookies or similar tracking technologies.

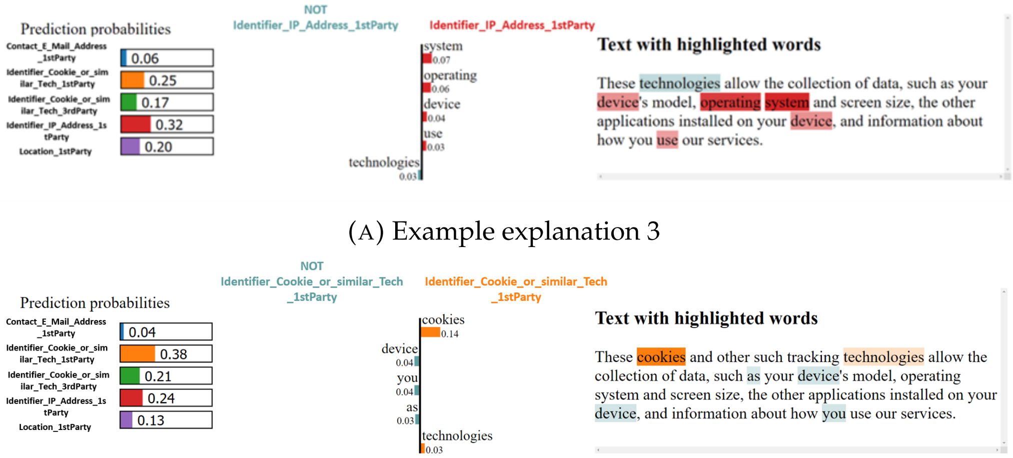
(A) Example explanation 2

Prediction probabilities
Contact_E_Mail_Address_1stParty 0.25
Identifier_Cookie_or_similar_Tech_1stParty 0.23
Identifier_Cookie_or_similar_Tech_3rdParty 0.23
Identifier_IP_Address_1stParty 0.13
Location_1stParty 0.15
NOT Contact_E_Mail_Address_1stParty
Contact_E_Mail_Address_1stParty
cookies 0.09
number 0.06
phone 0.05
use 0.04
us 0.04
Text with highlighted words

As explained above, you may either volunteer to us certain information (such as your phone number), or we may automatically collect certain information, such as through the use of your mobile device system's permissions, or through the use of cookies or similar tracking technologies.

(B) Counterfactual explanation 2

FIGURE 4.5: Example and counterfactual explanation 2


Prediction probabilities
Contact_E_Mail_Address_1stParty 0.06
Identifier_Cookie_or_similar_Tech_1stParty 0.25
Identifier_Cookie_or_similar_Tech_3rdParty 0.17
Identifier_IP_Address_1stParty 0.32
Location_1stParty 0.20
NOT Identifier_IP_Address_1stParty
Identifier_IP_Address_1stParty
system 0.07
operating 0.06
device 0.04
use 0.03
technologies 0.03
Prediction probabilities
Contact_E_Mail_Address_1stParty 0.04
Identifier_Cookie_or_similar_Tech_1stParty 0.38
Identifier_Cookie_or_similar_Tech_3rdParty 0.21
Identifier_IP_Address_1stParty 0.24
Location_1stParty 0.13
NOT Identifier_Cookie_or_similar_Tech_1stParty
Identifier_Cookie_or_similar_Tech_1stParty
cookies 0.14
device 0.04
you 0.04
as 0.03
technologies 0.03

Text with highlighted words

These technologies allow the collection of data, such as your device's model, operating system and screen size, the other applications installed on your device, and information about how you use our services.

(A) Example explanation 3

Text with highlighted words

These cookies and other such tracking technologies allow the collection of data, such as your device's model, operating system and screen size, the other applications installed on your device, and information about how you use our services.

(B) Counterfactual explanation 3

FIGURE 4.6: Example and counterfactual explanation 3

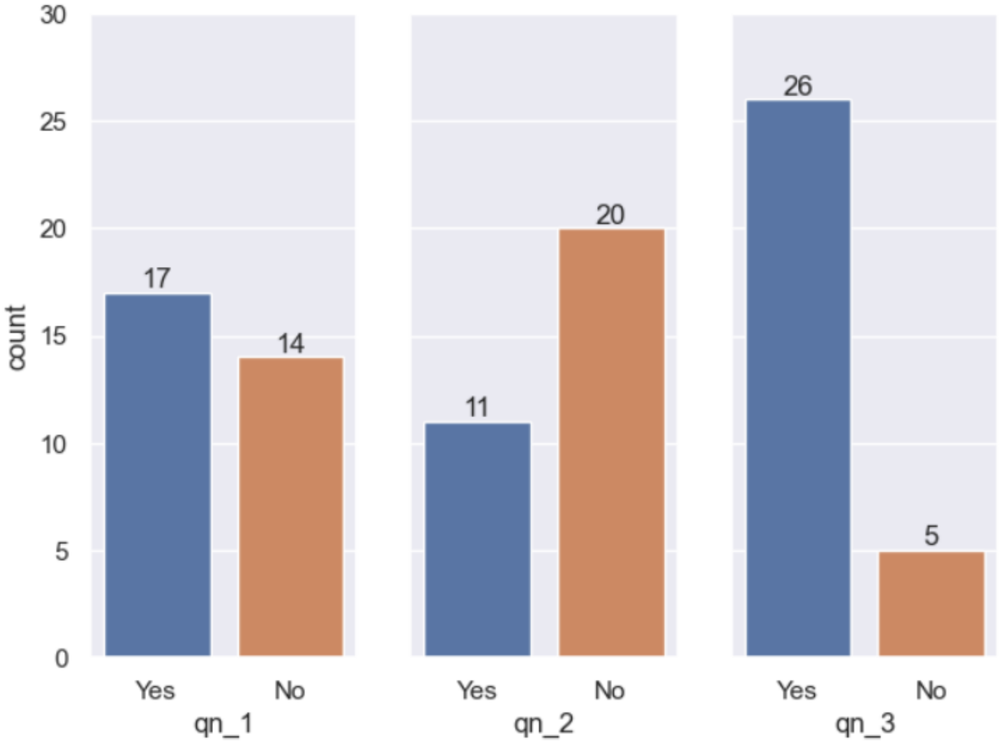
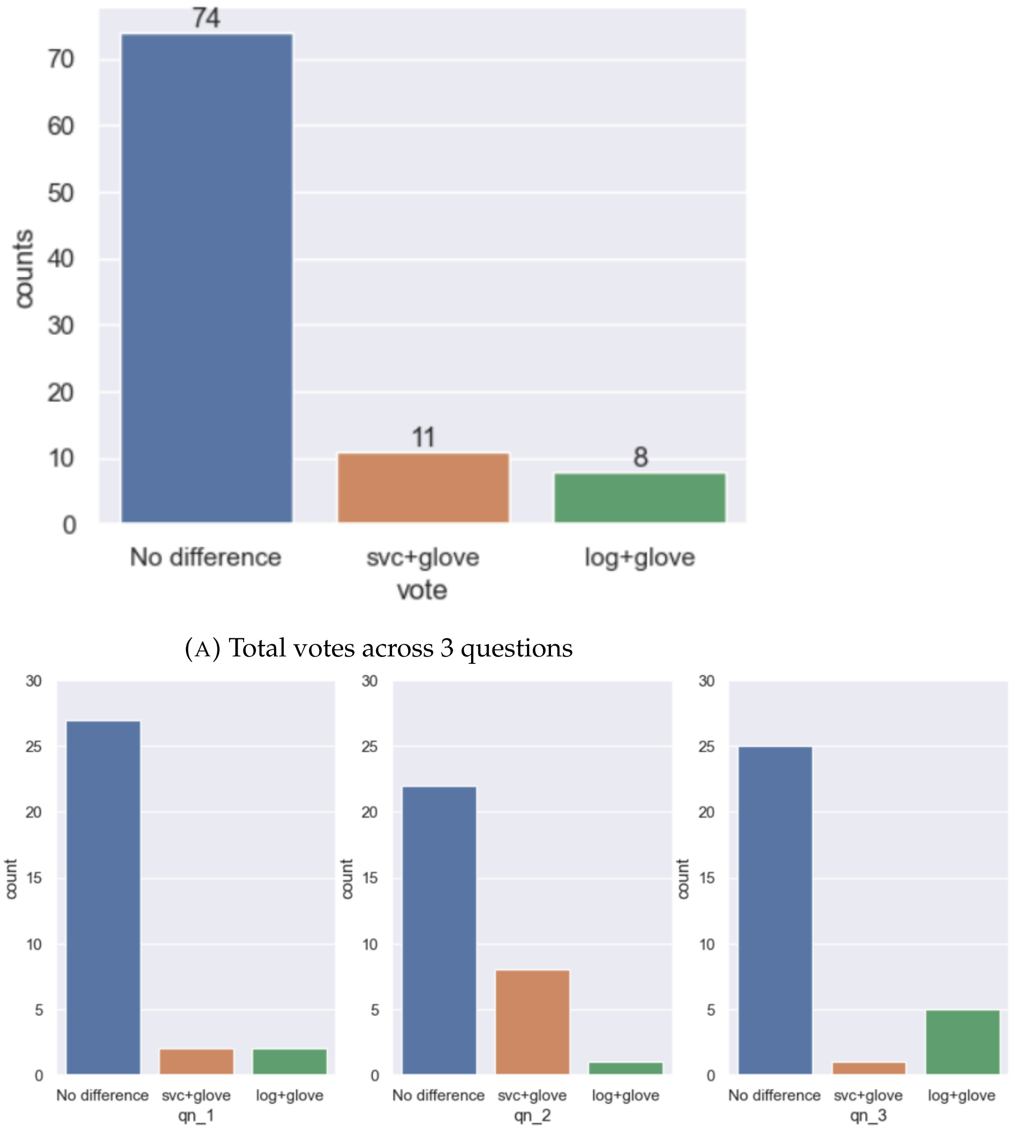


FIGURE 4.7: Votes for predicting whether counterfactual would be classified as `Identifier_Cookie_or_Similar_Tech_1st_Party`

4.1.4 Part 4 & 5: Testing which model and word representation is more explainable

As there were three questions in each section to test the explainability of each pair, the maximum total number of votes an option could get was $31 \times 3 = 93$. I totalled up the votes for each option for each part and each question in Figure 4.8 and 4.9.

Overall, respondents found no difference in explainability between logistic regression and SVC (Figure 4.8), while it was more contentious when comparing word representations, with a third split across the three options (Figure 4.9). Remember that the model performance for each pair actually did not differ significantly, as mentioned in Chapter 4. In terms of comparing respondents' votes with the "ground truth" metric of model performance, we would expect most respondents to note that there are no differences between the explainabilty of the model. This was seen when comparing logistic regression vs SVC but not seen in the case of comparing Tf-IDF and GloVe. Votes for word representations had a higher spread of votes across the three options. While there is no majority consensus, the fact that the votes were almost equally split instead of being heavily weighted in favour of one category shows that respondents were more undecided in the explainability 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 explainability even when comparing models of relatively same performance, as picked up by the differences in distribution of the respondents' votes.



(A) Total votes across 3 questions

(B) Votes per question

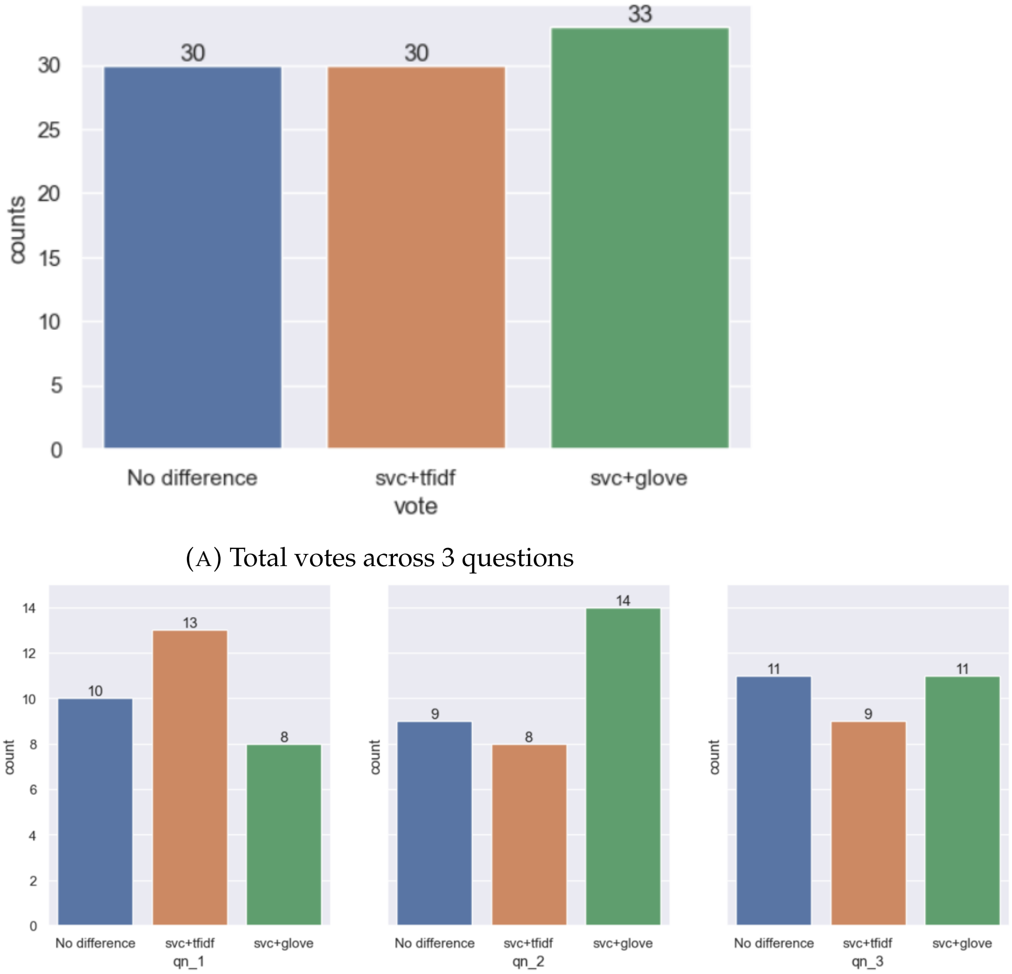
FIGURE 4.8: Testing for which model was more explainable: Respondents' votes to whether SVC + GloVe or Logistic regression + GloVe were more explaninable

4.1.5 General discussion of results

Here are some themes that arise from the preceding discussion:

Explainability is context sensitive.

Whether the self-reported explainability metrics significantly decreased



(A) Total votes across 3 questions

(B) Votes per question

FIGURE 4.9: Testing for which word representation was more explainable: Respondents' votes to whether SVC + TfIDF or SVC + GloVe were more explaninable

or increased after viewing the explanations depended on the type of metric and the context. Respondents likely placed different emphasis on different values depending on the contexts, which resulted in differing evaluations of the the same explanation. While this finding is uncontroversial, it confirms current XAI research, and also demonstrates the complication of developing "more explainable" models when evaluating such explainability is context dependent, compared to typical metrics of performance such as precision / recall / F1. Further, perceptions of explainability can also be influenced by pre-existing expectations that users of models have. If users have the false conception that anything produced by a computer is "objective" and "trustworthy" without being aware that machine learning is inherently fallible, they are likely to report a reduction in explainability after viewing explanations that expose the fallibility of these models.

Global and local explainability should be evaluated in tandem.

Global explanability can decrease even though local explanability may not be an issue. This is one of the limitations of local explainability techniques, as it allows users to understand the model through specific examples. Hence, explainability is highly dependent on the types and number of examples shown to users. If the examples chosen are insufficient or demonstrates an inuitively inconsistent pattern, users may be more confused about the model as a whole which affects global explainability. So it is important to consider global explanability by using different XAI techniques beyond just local explanations. Nevertheless, not all use cases require global explainability, such as perhaps the context of the consumer who may only be concerned with his own instance.

Models with similar performance metrics can have different perceptions of explainability.

Different models and text representations can have different perceptions of explanability, such as the instance when Tf-IDF is compared against GloVe embeddings. While this finding might not surprising given that Tf-IDF is a purely count based metric that does not take into account word context unlike GloVe, such "explanability characteristics" of models can be taken into account when explanability is important for a particular use case beyond simply evaluating based on traditional performance metrics. Nevertheless, it is also important to note that the results obtained here are restricted to the particular dataset and implementation used, and I am not asserting that a particular model or text representation has an "intrinsic explainability quotient".

5 Conclusion

5.1 Findings

We can now answer the research questions that were posed in Section **??**:

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

   Overall, SVC + Tf-IDF performed the best across the 5 data practices according to the weighted F1. This corresponds with the findings of Zimmeck et al. Particularly when classifying `Identifier Cookie 1st Party`, both SVC + Tf-IDF and logistic regression + Tf-IDF performed similarly with the same F1 score. It is surprising that GloVe embeddings performed worse compared to Tf-IDF, given that GloVe accounts for context as compared to Tf-IDF which is purely count-based.

2. Which classifiers are the most explainable 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 explainability. When Tf-IDF was compared to GloVe, respondents were undecided with an almost equal one-third split across the the three options (i.e. no difference, Tf-IDF & GloVe). So just by looking at these votes, respondents did not find any difference in explainability when comparing SVC to logistic regression, but were undecided whether Tf-IDF or GloVe was more explainable. Comparing the respondents' votes with performance metrics, there seems to be a slight (but non-statistical) relationship between explainability 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. Compared to logistic regression and SVC where latter did not consistently outperform the former, 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 negative trend of respondents' self-reported scores of interpretability and understanding for LIME. Hence, one possible interpretation is that the individual explanations got less explainable for each question. Another interpretation is that global explainability decreased because respondents were exposed to contradictory information about the model with each new explanation they viewed, rather than local explainability being affected. In terms of predicting counterfactuals, it was not possible to perceive any consistent trend given that the classifiers' predicted probabilities for some classes were almost equal for some of the counterfactual sentences.

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 (as the regulatory authority for data privacy) after viewing the classifiers' explanations?

   Effectiveness and trust significantly decreased for context 2 (PDPC), and fairness for context 3 (consumer). Only effectiveness significantly increased for context 1 (app developer). This could suggest that respondents believe that as a regulatory body, PDPC should retain more discretion in making decisions instead of relying on AI, even though using these models could increase automation. Respondents believe that the risks of making a wrong decision outweighs the possible benefits to efficiency. Having realised the fallibility of the model, respondents have reduced trust as well. Respondents as consumers also seem to be only concerned about their own benefits that result from wrong / correct predictions of the model. Hence, it is not "fair" that they would not be compensated for violations to their data privacy just because the model can make wrong predictions. Respondents as app developers, unlike PDPC, believe that wrong predictions are an acceptable risk when analysing privacy policies can be automated without needing to pay for external legal advice.

   Further, explainability metrics overall significantly decreased for context 2 and 3, which could suggest that respondents felt the risks of making a wrong prediction because of the model's fallibility outweighed any efficiency, or cost gained from automation from the perspective of PDPC or a consumer. Trust also significantly decreased overall for all contexts. This could suggest that the model's fallibility influenced respondents' trust the most across the contexts, which points to a possible mismatch of expectations of what respondents thought the model could do and what was actually the case.

5.2 Limitations and future work

There are a few limitations:

1. *Limited number of respondents ($n = 31$), diversity of expertise and age*: The intent was to conduct a more in-depth cross sectional study of the self-reported scores of the respondents according to their area and depth of expertise (measured by their major or current occupation) and their beliefs relating to AI and data privacy. This analysis could be used to support certain inferences made earlier. For example, the inference was made that explainability metrics significantly decreased for the contexts of PDPC and consumer but not for the app developer because respondents believed that the app developer should have enough programming expertise to understand the classifier and does not need to rely only on the visualisations provided. To support this claim, a significance test could be used on the scores reported by respondents with data science background to check whether there was a significant decrease, and compared with those respondents without a data science background. 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 these more in-depth analysis.

2. *Analysis is limited to the context of the APP-350 corpus, classifiers used, and LIME:* As mentioned above, assessing XAI is highly context dependent, and therefore the foregoing analysis should also be seen in light of these particular components. For example, a finding was that respondents found no difference in explainability between SVC and logistic regression. However, this finding is restricted to this particular dataset, as these models and LIME would probably process another sentence from another dataset differently. Future work to evaluate XAI within the context of data privacy could include using another data privacy dataset, or other LIME implementations with different styles of visualisation.

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 and the order of these sentences shown to respondents were intended to show the limitations of the classifier. For example, the first two visualisations showed that "cookie" was the most important feature contributing to the classification, but the next two still classified the sentence as `Identifier Cookie 1st Party` even though "cookie" was not present. This paints a more confusing and inconsistent picture

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

of the classifier's logic, as compared to four examples consistently showing that "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 (Górski and Ramakrishna, 2021). Hence, if more consistent examples were chosen instead, respondents might have reported differently. If post-hoc techniques are used, future work could include an interactive study where respondents can choose to generate their own sentences for classification, which allows them to test whether 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 (Rosenfeld, 2021). 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.

4. *LinearSVC and logistic regression can be similar in their optimisation technique, which may influence explainability.*

5.3 Final comments

(Hacker and Passoth, 2022) - about intersection of law and explainability in different areas of law

(Yalcin et al., 2022) - reason for why efficiency and trust is emphasised in this capstone

(Vilone and Longo, 2021) - about how there are so many different metrics as part of explainability, including efficiency, trust, transparency, understandability etc.

(Kästner et al., 2021) - about the relationship between trust and explainability - can be used to explain survey results

(Waltl and Vogl, 2018)- Possible to assign metrics of explainability to AI models (like a rating)

Bibliography

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

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

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

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

— (2021b). "We, The Robots: Regulating Artificial Intelligence and the Limits of the Law". In: Cambridge University Press. Chap. 3. Opacity.

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

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

Dunning, David (2011). "Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance". In: ed. by James M. Olson and Mark P. Zanna. Vol. 44. Advances in Experimental Social Psychology. Academic Press, pp. 247–296. DOI: https://doi.org/10.1016/B978-0-12-385522-0.00005-6. URL: https://www.sciencedirect.com/science/article/pii/B9780123855220000056.

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

Górski, Łukasz and Shashishekar Ramakrishna (2021). "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, 60–68. ISBN: 9781450385268. DOI: 10.1145/3462757.3466145. URL: https://doi.org/10.1145/3462757.3466145.

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

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

Hacker, Philipp and Jan-Hendrik Passoth (2022). "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, pp. 343–373.

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

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

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

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

Katz, Daniel Martin et al. (2023). “GPT-4 Passes the Bar Exam”. In: *Available at SSRN 4389233*.

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

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

Mantelero, Alessandro (2014). “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, pp. 643–660.

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

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

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

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

Rosenfeld, Avi (2021). "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, 45–50. ISBN: 9781450383073.

SciPy (2023). *scipy.stats.wilcoxon*. Last Accessed: 2023-3-1. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html.

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

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

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

Wagner, Isabel (2022). “Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021”. In: *arXiv preprint arXiv:2201.08739*.

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

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

Zednik, Carlos (2021). “Solving the black box problem: A normative framework for explainable artificial intelligence”. In: *Philosophy & technology* 34.2, pp. 265–288.

Zimmeck, Sebastian et al. (2019). “MAPS: Scaling Privacy Compliance Analysis to a Million Apps”. In: *Proceedings on Privacy Enhancing Technologies* 2019.3, pp. 66–86.