



**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]

☐ Unrestricted access

Make the Thesis immediately available for worldwide access.

☒ Access restricted to Yale-NUS College for a limited period

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

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

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

Tristan Koh Ly Wey (Cendana College)

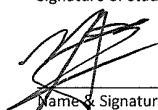
Name & Residential College of Student



Signature of Student

3 April 2023

Date



Michael Choi

Name & Signature of Supervisor

3 April 2023

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

Without God's abundant grace, this capstone and more would not have been possible. His love has been my constant source of strength in overcoming self-doubt to discover His fullness of joy.

I could not be more blessed to have my parents who have selflessly cared for me and seen me through the darkest of days, and to Law CF & 恩岭 for embracing me into the spiritual family and giving me another place to call home.

I owe a debt of gratitude to Prof. Choi & Prof. Chesterman, who have willingly taken on this strange mixture of MCS and law, and to Prof. Danvy & Prof. Francesca, because without their reassurance and nurture, I would never have dreamt of majoring in MCS.

I am privileged to have shared fond memories, laughter, academic and non-academic grumbling alike with Rayner, Nat, Ziyang, and Shaf, who have made this five-year sojourn all the more transformative. Special mention must go to the thankless toil of the cleaning and catering aunties and uncles, without whom this whole residential experience would not have lasted more than half a day.

Finally, I thank my survey respondents because without their participation, I would only have half a capstone. To you, my reader, I hope you find this capstone as insightful to read as I did writing it.

"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

Natural language processing (NLP) techniques have increased substantially in performance and can potentially automate some areas of legal reasoning and writing. However, increased performance has come at the cost of increased opacity of the models. XAI has been developed to combat this opacity. Nevertheless, it is unclear how such explanations should be evaluated and whether these explanations are effective because effectiveness depends on the varying purposes and end-users of the explanations. Separately, the stakeholders of data privacy regulation (i.e. organisations, data privacy regulatory authorities, and consumers of data services) can benefit particularly from the use of XAI. Hence, this capstone fills a gap in research by evaluating the effectiveness of XAI within the context of data privacy.

This capstone trains and reports the performance of machine learning classifiers that detect data practices in PlayStore apps' data privacy policies. Then, an XAI technique, LIME, is applied on these models to produce explanations of classifications of selected data practices. Finally, the

explanations are evaluated by human evaluation using a survey which uniquely utilises varied types of questions to measure different values linked to explainability (effectiveness, risk, fairness and trust) from the perspective of different stakeholders of data privacy. In terms of model performance, I achieve a 5-class weighted average F1 score of 66% using SVC with Tf-IDF. In terms of human evaluation, after viewing the explanations, respondents reported a statistically significant decrease in overall trust and understandability of the models, amongst other more specific findings.

Keywords: eXplainable AI, human evaluation of XAI, natural language processing, machine learning, data science, data privacy

Word count: 12,000

“Any sufficiently advanced technology is indistinguishable from magic.”

Arthur C. Clarke

Contents

Acknowledgements	ii
Abstract	iii
1 Introduction	1
1.1 Motivation and significance	1
1.2 Research questions and roadmap	7
2 Literature review and methodology	10
2.1 Dataset: The APP-350 Corpus	10
2.2 Classifiers	12
2.2.1 Literature review	12
2.2.2 Methodology	14
2.3 XAI methods for NLP	17
2.3.1 Literature review	17
2.3.2 Methodology	18
2.4 Evaluating the effectiveness of XAI methods	20
2.4.1 Literature review	20
2.4.2 Methodology	23
3 EDA and classifier performance	28
3.1 EDA	28

3.2	Performance of classifiers for top N practices	29
3.3	Performance of individual classifiers for top 5 practices . .	33
4	Discussion of survey results	37
4.1	Part 2 & Part 6: Comparison of self-reported scores of explainability across the three contexts	37
4.2	Part 3: Testing whether viewing more visualisations affected explainability	48
4.2.1	Analysis of reported interpretability and understandability	48
4.2.2	Analysis of predicting counterfactuals	51
4.3	Part 4 & 5: Testing which model and text representation is more understandable	53
5	Conclusion	56
5.1	Findings	56
5.2	Limitations and future work	58
	Bibliography	61

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?" [2]. This question seems close to getting an answer in 2023 with the release of ChatGPT [22] to the general public. Consider the following response from ChatGPT:

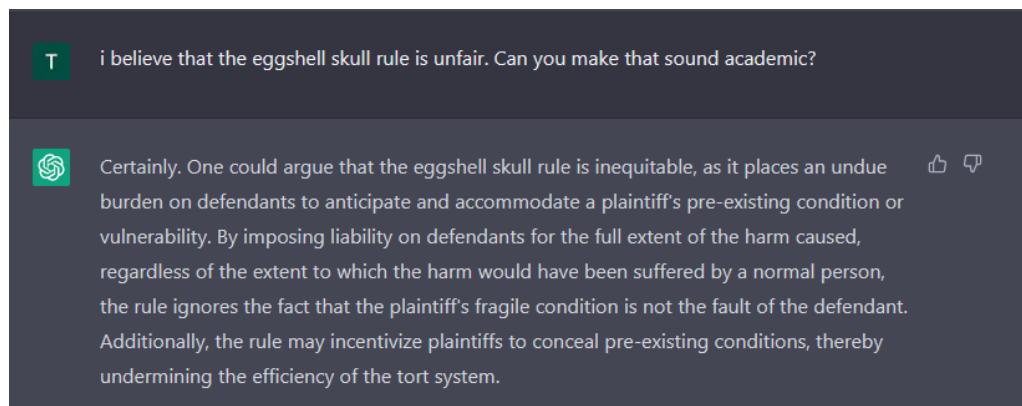


FIGURE 1.1: Conversation with ChatGPT about the eggshell skull rule.

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

ChatGPT and GPT-4 are the most recent AI models that uses advanced natural language processing (NLP). NLP is a branch of AI that gives computers the ability to understand text and spoken words as how human beings would understand [16]. While NLP techniques have substantially increased in performance in recent years, it has come at the cost of the explainability of their predictions because of models that are architecturally more complex and opaque [8]. This issue of explainability is exemplified by the last question I posed to ChatGPT about its view on the egg-shell skull rule:

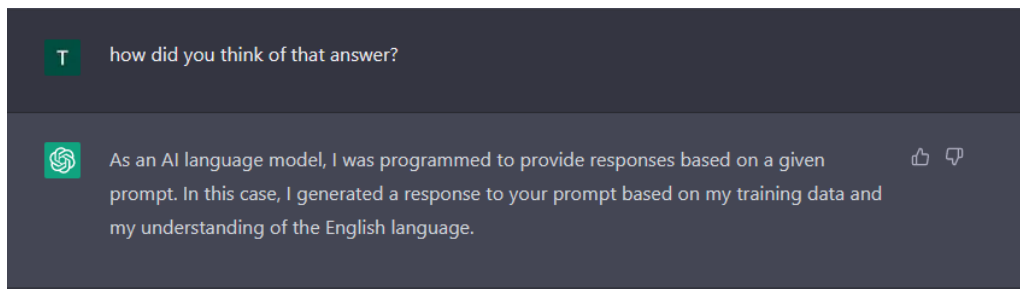


FIGURE 1.2: Conversation with ChatGPT about the egg-shell skull rule.

ChatGPT does not seem to be able to explain its views like how a typical human would. Further prompting led ChatGPT to provide a list of academic papers that support its view. However, there are issues of hallucination, where ChatGPT generates answers about facts that are false or non-existent [1]. This lack of explainability (and potential inaccuracies of the explanations) could significantly hinder 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 client¹.

¹According to r5(2)(c) of the Legal Profession (Professional Conduct) Rules 2015.

A lawyer that relies on the analysis of legal tech tools but does not understand how the analysis was produced could be considered lacking in diligence and competence.

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 proficient in data science to the extent required to understand the logic of machine learning models without aid. Thus, eXplainable AI (XAI) techniques and research have been rising in popularity and has most recently found its way into the the EU's proposed AI Act that places explainability requirements on AI providers [13]. XAI refers to methods that attempt to help the end user of a model understand how it arrives at its result [6], which reduces the model's opacity. XAI can therefore bridge the gap between the lawyer and the data scientist by using XAI techniques to explain the predictions of AI models used in legal decision making.

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" [6]. Interpretability is sometimes used to describe the model's internal logic and explainability as the ability of the user to understand that logic. However, explainability is also described as techniques to explain the model's logic post-hoc without necessarily being representative of the model's true decision [26]. Though the AI Act frequently refers to different levels of transparency, there is no explicit definition of transparency. Transparency

could involve mandatory disclosure of an AI system used in decision-making which is a much broader concept than transparency used in current XAI literature that covers only the algorithmic process [13]. In this capstone, I use explainability, interpretability, understandability and transparency interchangeably to refer to the overall goal of XAI as stated above.

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 understandable, laypersons might be more confused if too many details are provided [26]. 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 [12]. 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 following 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 day² [32]. 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 [21]. 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 processing³. 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, using opaque AI could affect

²Assuming that a consumer visits 1462 unique websites each year.

³This is a plausible interpretation when reading Recital 71 in conjunction with Art 22 of the GDPR.

the organisation's compliance of 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 used in decision making to build understanding and trust with those that use their products [24]. In terms of communication policies, organisations are advised to develop guidelines on the type of explanations and when to provide them. Despite debate about the scope of the right to explanation and whether it is binding [4], 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.

One of the state's concerns when AI is used in legal decision making is the integrity of the judicial system and regulatory activities since the legal system depends as much on the justification of its decisions as it does on the decisions themselves [5]. 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 decisions⁴. When overturning the PDPC's decisions, the higher courts do not disagree with the outcome but they disagree with the reasoning of the PDPC's decision. Assuming that AI models are sophisticated enough to write (or at least assist the writing) court judgements, if there was no explanation of the model's writing, there would be little basis for the higher courts to overturn judgements made by lower courts.

Hence, the importance of XAI applies across all three stakeholders in

⁴Per s48Q, s48R and s54 of the PDPA.

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. 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.

This capstone's objectives come broadly within the evaluation of XAI techniques. While there have been studies that evaluate XAI through human evaluation [30] and non-empirical research that investigates the relevant standard for explainability across different areas of law [14], I adopt an uniquely empirical methodology to assess XAI within a data privacy context across the different use cases of stakeholders. The closest study using similar methodology is one that asked lawyers to rate visualisations generated by different XAI methods [10]. However, this study was conducted using a dataset containing cases from the US Board of Veterans' Appeals, and did not consider evaluating how explainability can vary with different purposes of the explanations. To this end, this capstone also serves as a trial for an unique **survey design** that varies the purpose of explanations to assess how explainability and values related to explainability change.

1.2 Research questions and roadmap

I have four research questions⁵:

⁵All code and analysis can be found [here](#).

1. Which AI models ("classifiers")⁶ perform the best on a data privacy dataset in terms of traditional performance metrics? (Section 3.3)
2. Which classifiers are the most understandable to users that include laypersons and users with domain knowledge in data science and the law? (Section 4.3)
3. How would users rate the explainability of a selected XAI technique? (Section 4.2.1)
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 and the PDPC? (Section 4.1)

To investigate these questions, I train and evaluate the performance of classifiers that classify sentences in apps' data privacy policies to a particular data practice. A privacy practice is a section of an app data privacy policy that describes a certain behaviour of an app which has privacy implications⁷. XAI methods are then used to visualise why the models made such predictions. These visualisations are assessed for effectiveness by conducting a survey asking respondents to rate these visualisations according to certain metrics and values related to explainability and by using different types of questions.

Chapter 2 introduces the dataset and provides a literature review and justification for the methodology used in this capstone with regard to the NLP classifiers, XAI techniques, and human evaluation of XAI. Chapter 3

⁶I refer to AI models as classifiers from hereon as the task that the AI models perform in this capstone is classifying sentences to a data practice. I elaborate more on this distinction in Section 2.2.

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

presents an exploratory data analysis (EDA) of the dataset and reports classifier performance. Chapter 4 discusses the survey results and Chapter 5 concludes the capstone.

2 Literature review and methodology

There are four major components to this capstone: The dataset, classifier training, application of XAI techniques to the trained classifiers 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 classifiers of sufficient baseline performance, I focused

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

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

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

TABLE 2.2: List of modalities.

on training classifiers on the top 5 practices by frequency¹. 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 project to train models to conduct a privacy census of 1,035,853 Android apps [34]. The authors downloaded the data privacy policies of apps from the Play Store with more than 350 million installs and 103 randomly selected apps with 5 million installs. In total, there were data privacy policies of 350 apps. All 350 policies were annotated by one of the authors, a lawyer with experience in data privacy law. To ensure reliability of annotations, 2 other law students were hired to double annotate 10% of the corpus. With a mean of Krippendorff's $\alpha = 0.78^2$, the agreement between the annotations exceeded previous similar research.

Since the focus of this capstone is to assess XAI methods within a data

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

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

privacy context, the APP-350 corpus was chosen as it contains real-world data privacy practices. Training classifiers 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 easier validation of classifier performance.

The annotated privacy policies were originally in .yaml, with one .yaml file containing one app data privacy policy. Each data privacy policy is labelled at both the sentence and segment³ level, but to limit the scope of this capstone I focus only on sentence level. The data was restructured from .yaml to .csv.

2.2 Classifiers

2.2.1 Literature review

In NLP, there are two components of a text classifier: The text representation, and the model. In this capstone, I distinguish between "model" and "classifier". "Model" refers to the machine learning algorithm used to generate predictions, while "classifier" refers to the combination of a text representation and model that classifies a segment of text to a particular category. I make this distinction because part of this capstone's methodology is testing the explainability of the different combinations of models and text representations, of which both components are needed to form a classifier. However, for the **survey design**, I use "model" to refer to the entire classifier since the respondents did not need to know the distinction.

³A segment is similar to a paragraph.

Text representations are methods that translate plain text to mathematical representations that can be understood by a computer. The simplest text representation is a count-based approach such as Bag-of-Words, where each word is represented by its frequency in the sentence. These can be categorised as sparse vector representations. However, they do not capture the semantic and syntactic differences of words [20]. 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 [20]. However, to capture such semantic meaning requires more abstraction, increases the opacity and decreases the interpretability of NLP models. Hence, there is an inverse relationship between performance 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 [33]. In generating text representations of the APP-350 corpus, [34] 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 of the models. 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, [34] used support vector machines for classification (SVCs) to achieve reasonable performance. Individual models were then trained for every data practice. For all the classifications (except for four categories), they trained SVCs with a linear kernel and five-fold cross validation.

For the top 5 data practices that this capstone is concerned about, [34] 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 practices⁴.

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 networks⁵, I focus only on training classical models in this capstone. Since sparse vector representations and classical models have performed well for the APP-350 corpus 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

⁴Not all the data practices were used in training and testing these classifiers, so there is no reference performance for some of the data practices that are used in this capstone.

⁵For 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 document from others. The Tf-IDF metric for a word in a document is calculated by multiplying two different metrics:

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

In comparison, GloVe as a dense vector representation is able to better model semantic differences mathematically in the following way: King – Man + Woman = Queen [31]. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The length-50 vectors used in this capstone are pre-trained using this algorithm on Wikipedia and news articles [23]. The main idea behind GloVe is to learn text representations by considering the statistics of words that occur together in a corpus. These statistics capture how frequently different pairs of words occur together in the text. GloVe builds a matrix from these statistics, where each element represents the number of times two words appear together in the same context. The matrix is then factorized into a product of two low-rank matrices, where each row in one of the matrices represents a vector representation of a word. This results in dense, fixed-size vector representations for each word in the corpus that encode their meaning and relationships with other words in the corpus.

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

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

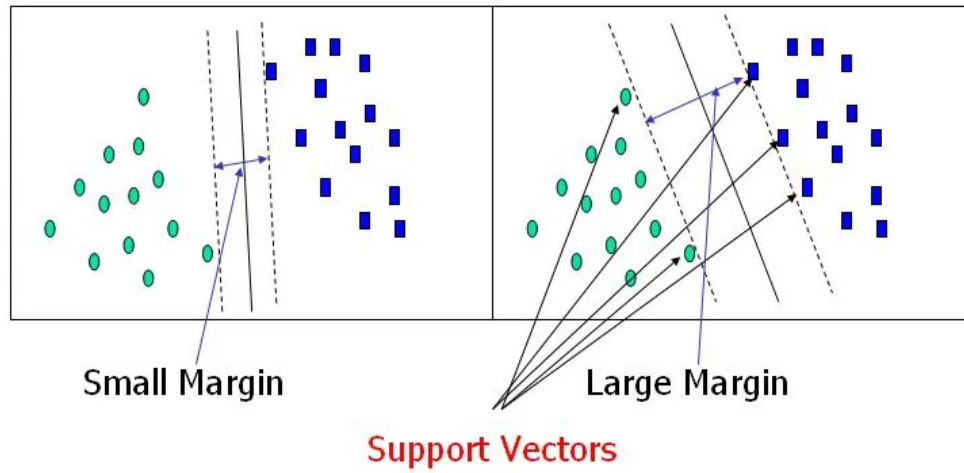


FIGURE 2.1: Finding the hyperplane in SVC [3].

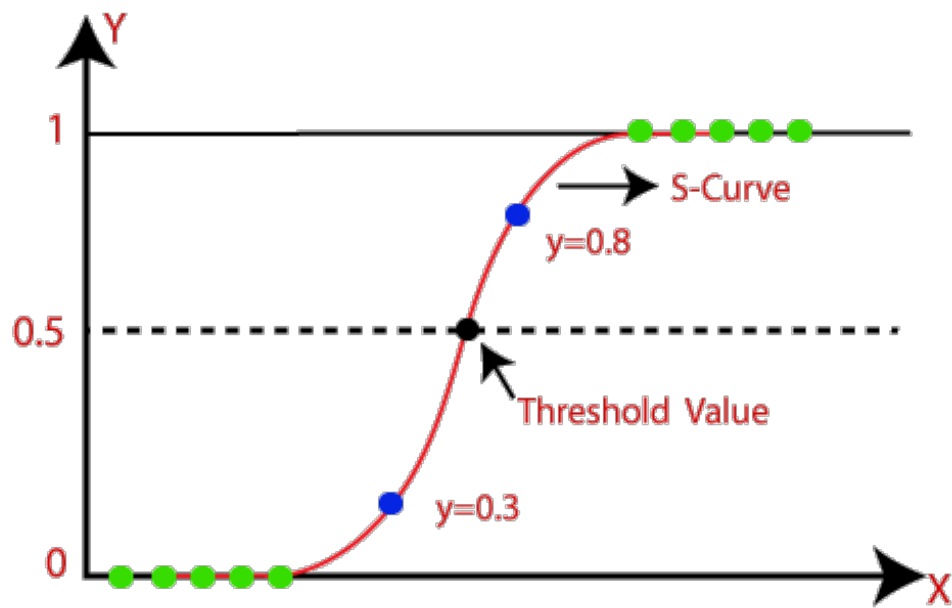


FIGURE 2.2: Sigmoid function in logistic regression [17].

2.3 XAI methods for NLP

2.3.1 Literature review

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

seen in Table 2.3.

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

TABLE 2.3: The four categories of XAI for NLP, adapted from [6].

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 addition 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 explainability have to be used instead.

2.3.2 Methodology

I use LIME (Local Interpretable Model-agnostic Explanations), a local post-hoc XAI method that is also model agnostic [25]. 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 [25].



FIGURE 2.3: Sample visualisation of LIME

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 techniques [30]. Standards of explainability are highly dependent on the purpose of generating the explanations. A few purposes have been suggested [7]:

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 [29] might not apply *mutatis mutandis* to the GDPR. Indeed, different standards for explainability have been argued to apply across contract, tort and banking law [14]. In any case, this capstone makes no claim as to the applicable standard for explainability for data privacy, and merely serves as empirical insight of explainability from the perspectives of the three stakeholders in data privacy.

In terms of methods of assessing XAI, human evaluation has been frequently utilised [6]. 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 [7].

Human evaluation has been conducted within a legal context [10]. 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 six correctly classified sentences on a scale from 0 (worst) to 10 (best). 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.

There is also 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 [26]. The relative importance of these values and overall requirement of explainability is necessarily dependent on the purpose and user of the explanation. Values such as fairness, bias, risk and trust could be argued to be of specific importance in a legal context which includes data privacy. The notion of "rule of law" encompasses these values and legal systems are arguably judged according to such values [11]. Nevertheless, there is no definitive position as to how these values can be used to evaluate XAI techniques that are specifically used within a legal context.

2.4.2 Methodology

With the foregoing discussion in mind, I expand upon the methodology of [10] by testing the effectiveness of LIME by varying the different purposes of explanations and methods of evaluation that were identified above. The primary method of evaluating LIME in this capstone is human evaluation through a quantitative⁶ survey by measuring respondents' agreement with questions on a 5-point Likert Scale. The survey was designed containing the following parts⁷:

1. **Part 1: Demographic data and beliefs relating to data privacy and AI.**

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

2. **Part 2: Rating of explainability metrics from the perspectives of three stakeholders in data privacy.**

⁶In Part 3 of the survey, I asked respondents to qualitatively explain why they provided the score they did of the understandability of the visualisation. However, due to time constraints, I did not use these responses in the analysis.

⁷The full survey can be found [here](#), and a summary of the responses can be found [here](#). Visualisations that were of significance to the analysis are provided later in Chapter 4.

After a brief explanation of the data practices and how NLP works, respondents were provided with three contexts (app developer, PDPC & user). All three contexts relate to solely relying on the predictions of the classifier to make a decision in relation to a finding of a breach of the PDPA, and are meant to vary the purposes of the explanation according to the three purposes suggested by [7] above.

The app developer context varies the programming expertise, unlike the PDPC which represents legal expertise, while the user has neither. The need for global explanations is incorporated in the PDPC context, since a regulatory authority would likely be assumed to be more concerned with systemic effects of using AI in their decision making because its decision making process affects how it decides other cases. In contrast, the app developer and user have more local needs for explanations, as they are concerned with one instance of reading a data privacy policy. The area and severity of incompleteness of the problem is increased for the PDPC, as using the classifier substitutes a substantial component of its legal reasoning process which is a specialised and specific skill. However, for the app developer and the user, there is a smaller area and severity of incompleteness since they are not equipped with legal reasoning in the first place. Using the classifier becomes more of a cost consideration between paying for human legal advice, or having the legal advice automated by the classifier.

The consequence of making a wrong decision was kept constant across the contexts by stating that a \$10,000 fine would be imposed if there was a finding that the PDPA was breached. The contexts

were also phrased such that the only legal requirement to find a breach of the PDPA was whether or not the classifier detected the presence or absence of a sentence that stated cookies were being used. Hence, the classifier completely substitutes the legal reasoning normally done by PDPC or a lawyer to reduce the legal knowledge required to understand the contexts.

After reading each context, the respondents were required to rate their beliefs relating to the use of the classifier in that particular context across four metrics ("explainability metrics"): Effectiveness of model, fairness, risk to society, and trust in the model. Respondents were told to use whichever definition of the metric they felt was the most appropriate. These metrics were chosen as they are values of importance specifically to a legal decision making context as mentioned above. The purpose is to investigate how these values that relate to explainability can vary according to the purposes of the explanation.

3. Part 3: Rating of LIME visualisations of four classifications by the same classifier (Logistic regression + Tf-IDF).

Respondents were asked to rate: (1) how far they understood why the model made the prediction, and (2) how far they found the visualisation easy to interpret. The intent is to evaluate both the underlying XAI technique (LIME) and the visualisation of the technique. For three classifications, respondents were given a counterfactual sentence replaced with words that were considered important by the classifier, and asked to predict whether the sentence

would be classified under Identifier Cookie 1st Party. Counterfactual simulation is meant to provide another method to measure explainability, aside from self-reported scores.

4. Part 4 & Part 5: Choosing the more interpretable visualisation from a pair of identically classified sentences.

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

5. Part 6: Rating of the same explainability metrics for the same three contexts.

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

The survey was hosted on Google Forms and respondents were anonymously recruited. 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 ethics requirements. No personal identifying information of the respondents was collected.

3 EDA and classifier performance

3.1 EDA

There are a total of 18,829 annotated sentences and the top 10 frequently occurring practices make up close to 60% of all sentences (Figure 3.1). The bottom 10 frequently occurring practices make up approximately 1% of the dataset. There does not seem to be much variation in sentence length for the top 10 frequently occurring practices, since the standard deviation of the mean is approximately 2.5 words and the median 1.9 words across the top 10 practices. This is also seen by similar interquartile range (IQR) (Figure 3.3c). The mean of the sentence lengths for all top 10 practices are slightly greater than the median, which suggest some outliers with very long sentence lengths (Figure 3.2) that cause a slightly right skew distribution (Figure 3.3a). The distributions are also unimodal, with the mode around 20. Similar sentence length could indicate similar sentence complexity across the practices.

With regard to top 10 frequently occurring 4-grams in the top 5 practices (Figure 3.4), both 4-grams of Cookie 1st Party and Cookie 3rd Party contain many instances of "cookies" across their top 10 occurring 4-grams. It is difficult to tell from the 4-grams alone whether the sentences 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

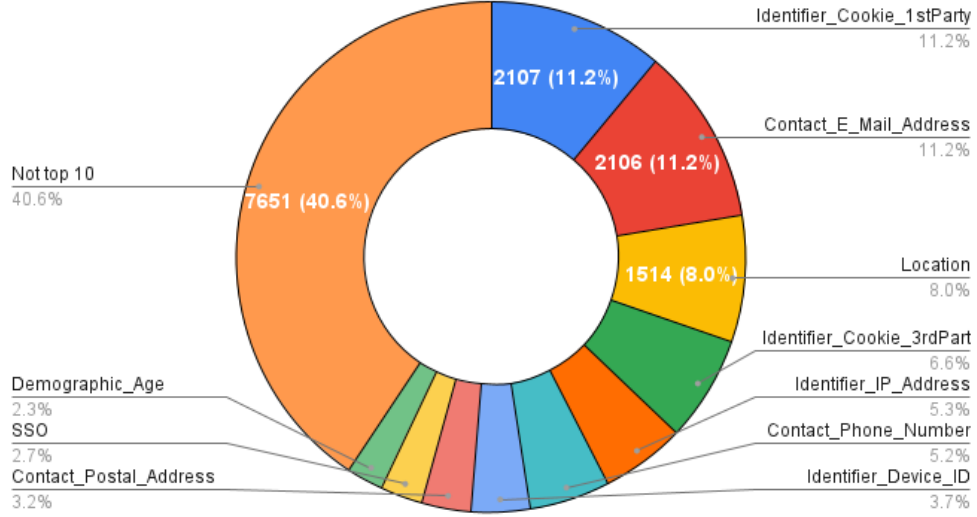
Data practices by proportion (n = 18,829)

FIGURE 3.1: Breakdown of data practices by proportion

both have instances of "IP address" and "mac address". Such similar 4-grams could affect the separability of the classes. In comparison, 4-grams of Email are clearly distinct from the rest of the data practices with unique tokens of "email" and "phone number".

3.2 Performance of classifiers for top N practices

I compare the weighted precision, recall and F1 scores of the classifiers below. Since there is neither a high risk of identifying false positives or 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. All classifiers were trained using one-vs-the-rest strategy for computational efficiency.

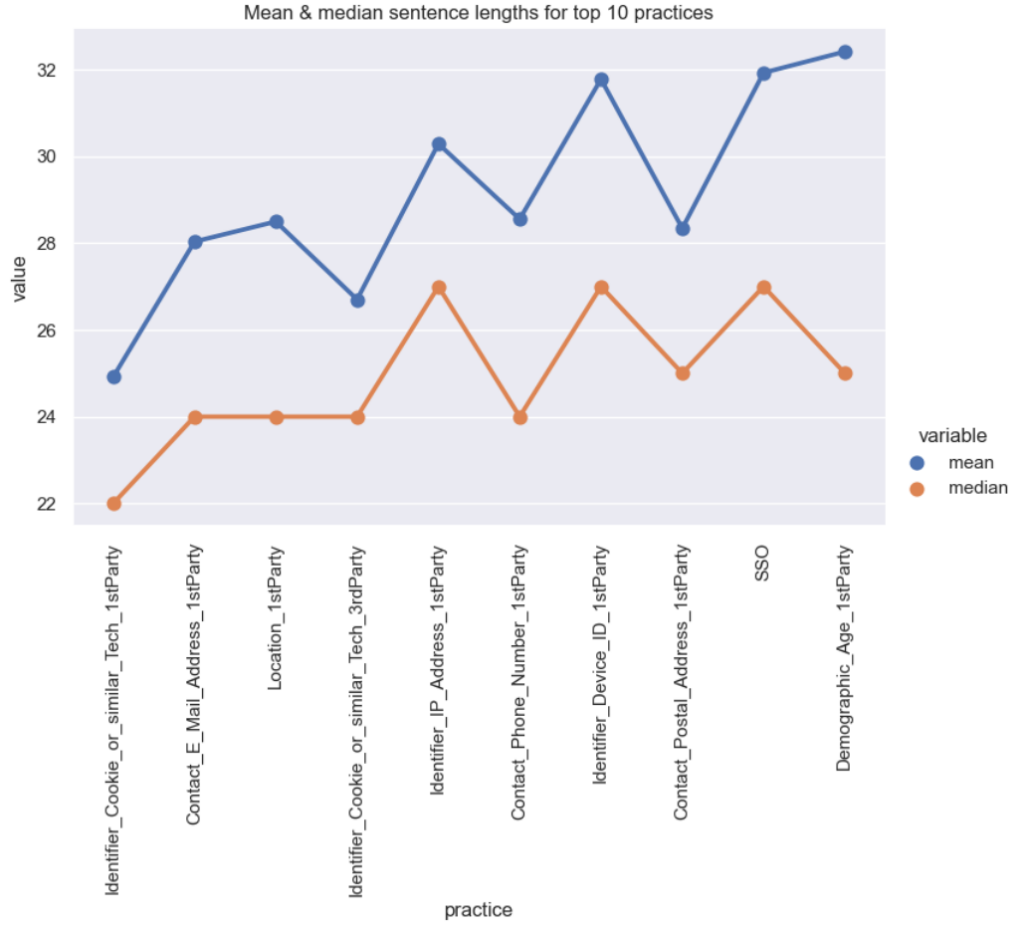
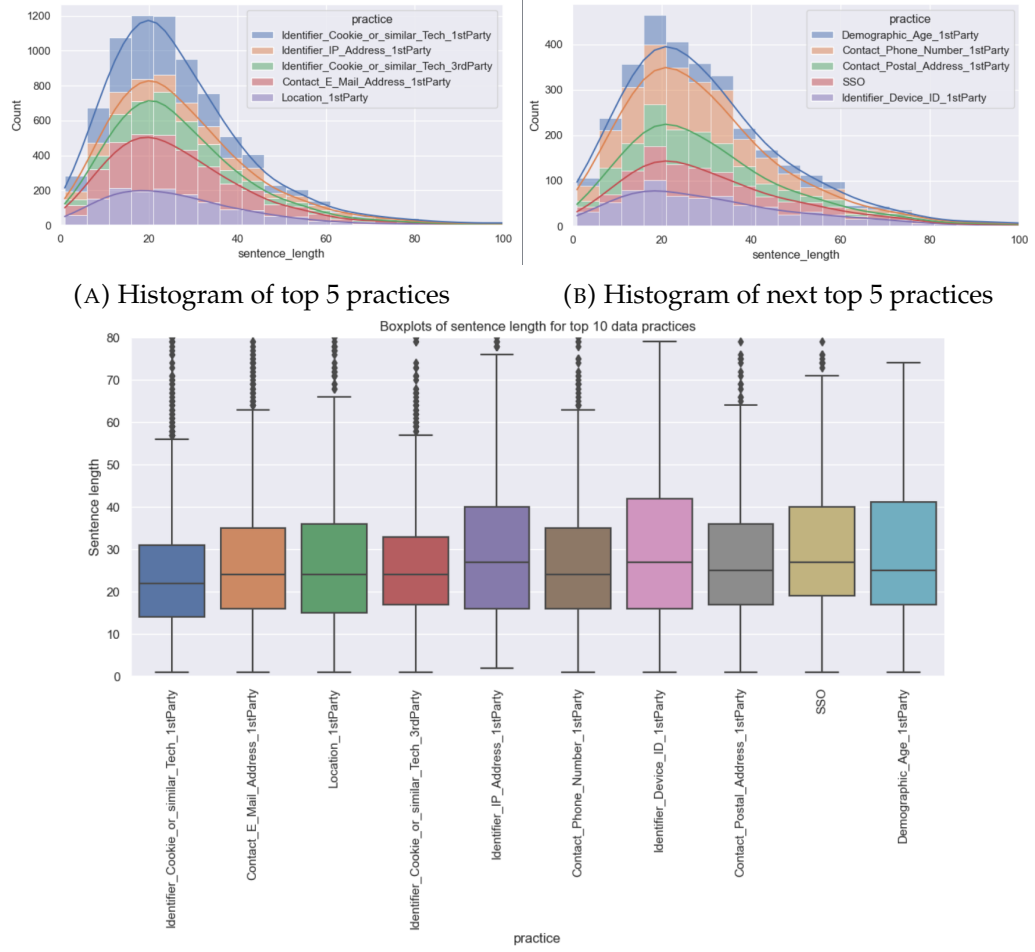


FIGURE 3.2: Mean and median sentence length

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 problem space for classification, 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 consistently performs the best across the metrics and across the top N frequently occurring practices (Figure 3.5). This corresponds with the findings by [34]. Also, performance of all the classifiers generally decreases with increasing N , which



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

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

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.

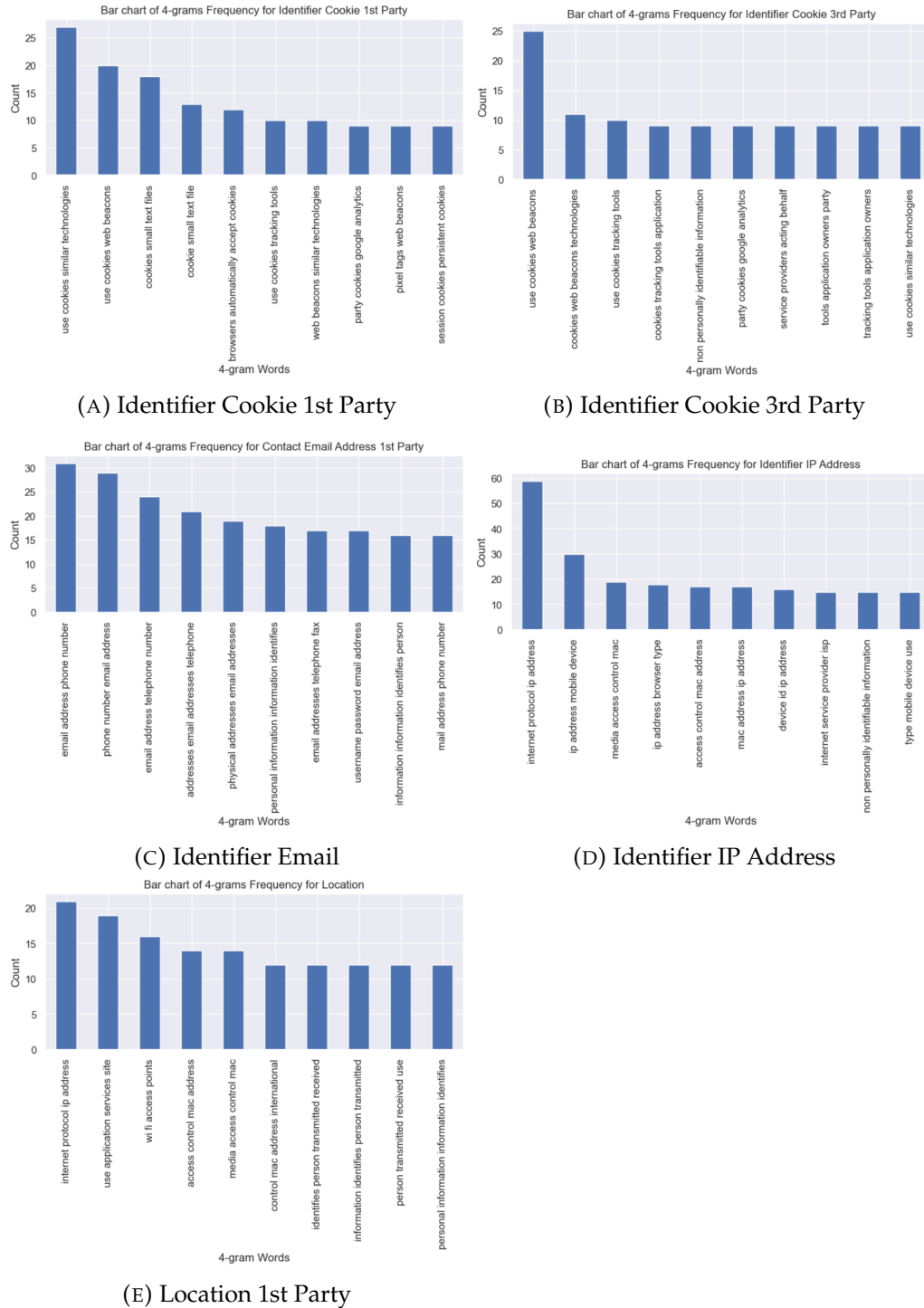


FIGURE 3.4: Top 10 most frequently occurring 4-grams in the top most frequently occurring 5 data practices at the sentence level

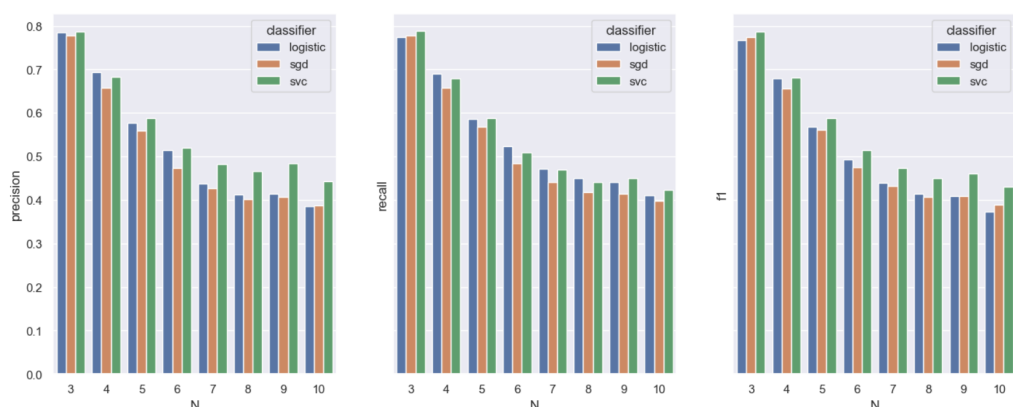


FIGURE 3.5: Weighted average P/R/F1 scores for top N ($3 \leq N \leq 10$) practices

3.3 Performance of individual classifiers for top 5 practices

I report the individual performance of all 4 combinations of text representations and models. With regard to P/R/F1 scores (Figure 3.6), similarly to what [34] found, SVC + Tf-IDF provided the best overall performance with the weighted average at 66%. This is a significant performance difference from the lowest performing model, logistic regression + GloVe at 54%. Specifically for `Cookie 1st Party`, both logistic regression + Tf-IDF and SVC + Tf-IDF performed similarly at 68% F1, but SVC + Tf-IDF had a more balanced performance across precision and recall. With regard to the ROC curves (Figure 3.7), looking at the micro-averaged AUC, the classifiers all performed similarly well within a small difference of 0.03. Similarly for `Cookie 1st Party`, the difference in AUC was insignificant at 0.02. Interestingly, 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 $0.91 - 0.85 = 0.06$.

In terms of performance for the individual classes, Email Address consistently performed the best across the classifiers, followed by Cookie 1st Party, Location, IP Address and Cookie 3rd Party. The difference in performance between the top performing class Email Address and Cookie 3rd Party can be up to 33% difference in weighted F1 depending on the classifier used. Specific class performance for Cookie 1st Party varied quite significantly from 68% using SVC + Tf-IDF, to 55% for SVC + GloVe. The performance differences of the classes might be accounted for by the uniqueness of the top 10 frequently occurring 4-grams as mentioned above. Since the 4-grams of Email were the most distinct compared to the rest, the classifiers could more easily separate Email from the rest of the classes which resulted in highest performance.

Overall, these performance figures confirm that SVC is the better performing model, though the performance difference between the benchmark logistic regression and SVC is not that significant (from 62% to 66% weighted average F1 when comparing between the classifiers that both use Tf-IDF). Further, the choice of text representation seemed to affect performance more than the choice of model. This is because 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 the semantic meaning of the tokens. These results correspond with a study measuring performance of supervised machine learning for classification tasks using

generalised text datasets [15], where LinearSVC followed by logistic regression were the highest performing models.

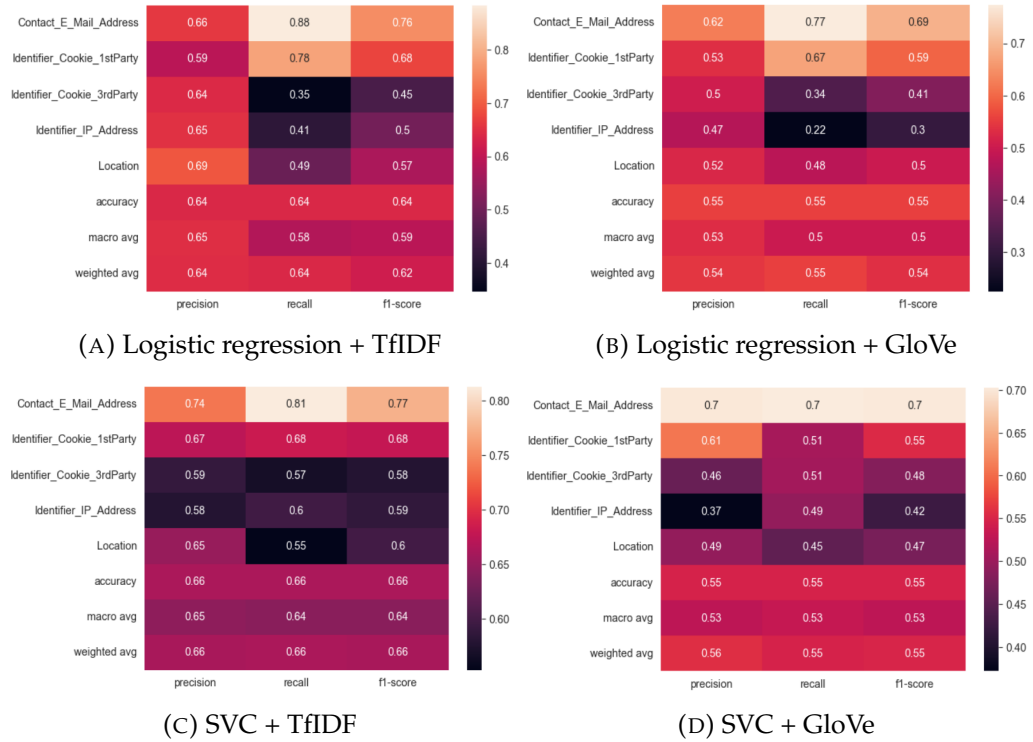
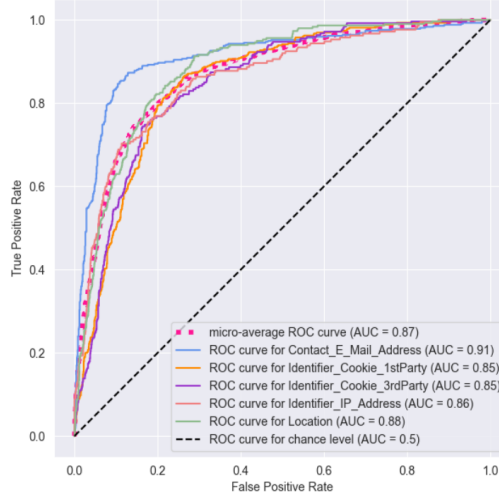
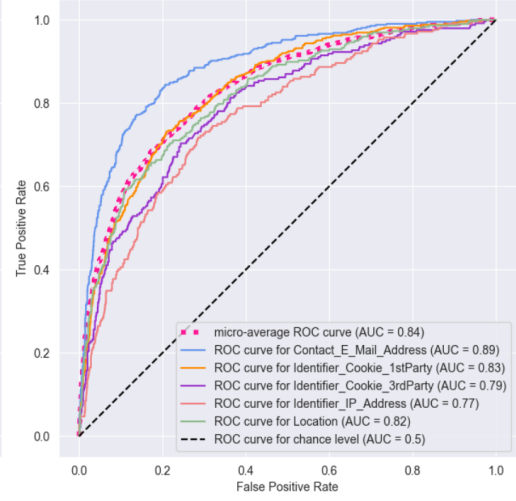


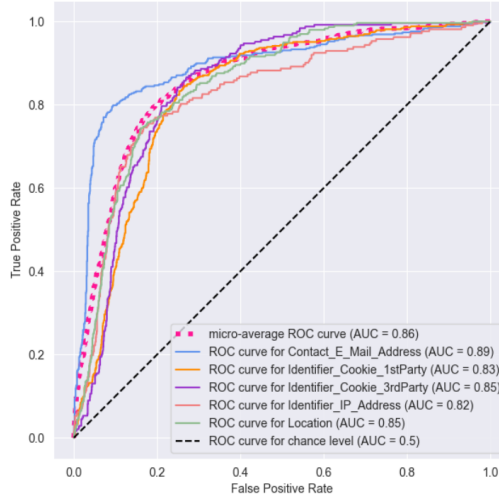
FIGURE 3.6: Performance heatmaps of the 4 classifiers (averaged over 5-fold cross validation)



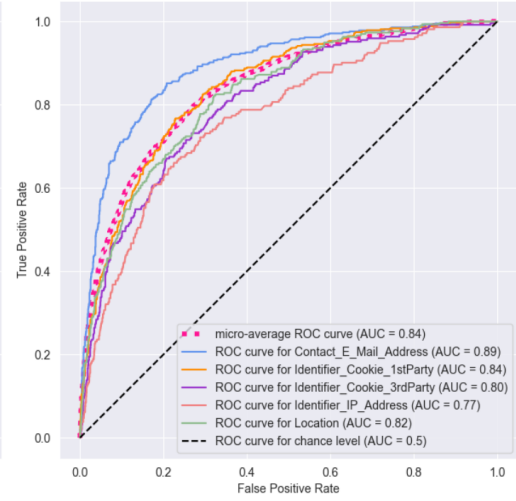
(A) Logistic regression + TfIDF



(B) Logistic regression + GloVe



(C) SVC + TfIDF



(D) SVC + GloVe

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

4 Discussion of survey results

In total, 31 responses were collected. The majority of respondents majored in law (58%) vs non-law (42%) respondents (Figure 4.1). Except for the questions relating to subject matter expertise (data privacy and AI), the level of agreement of law vs non-law respondents about their beliefs relating to data privacy and AI were about the same (Figure 4.2). Law respondents had less expertise in AI, while conversely, non-law respondents had less experience with data privacy.

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

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

Using the Wilcoxon Rank Sum Test as the distribution of the scores are likely nonparametric, I tested for the following, setting $\alpha = 0.1$:

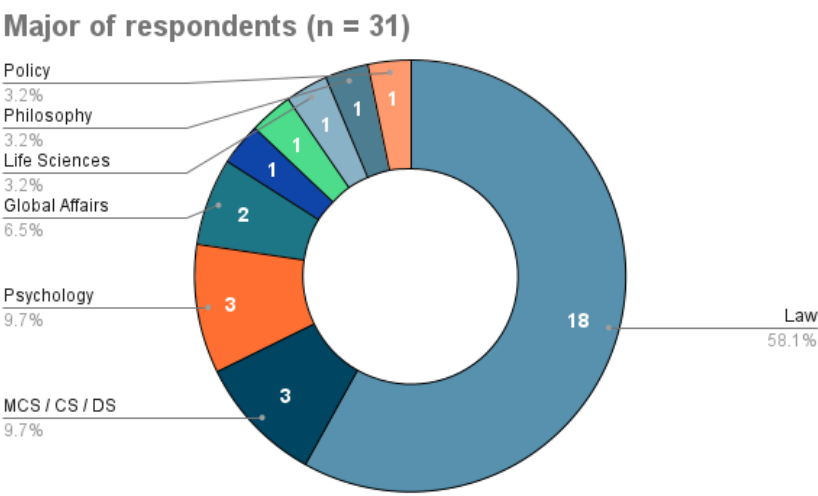


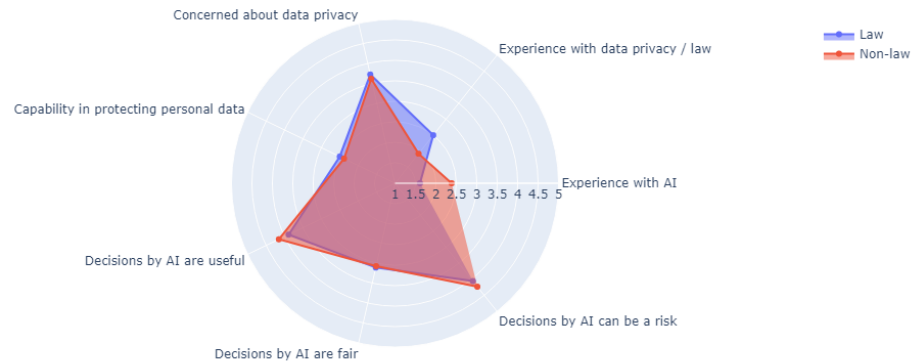
FIGURE 4.1: Breakdown of respondents' expertise by major

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

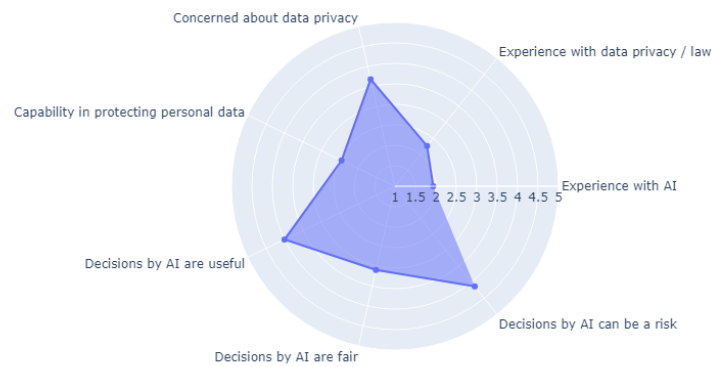
H1: There is an increase / decrease in scores after viewing the explanations.

The 1-sided test was used to check whether the distribution underlying the difference between the initial and final paired scores was symmetric below or above 0 [27]. Mathematically this difference can be stated as $d = i - f$, where i and f are the scores reported before and after viewing the explanations, and d is the difference. Hence, if $d < 0$, then $i < f$ and the scores increased after viewing. Conversely, if $d > 0$, then $i > f$ and the scores decreased after viewing.

p-values are reported in Table 4.1 and 4.2. For Table 4.2a and 4.2b, I took the mean of self-reported scores across metrics, and across contexts respectively. For Table 4.2c, the self-reported scores were averaged across



(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. $n = 31$)

both context and metric. Cells highlighted in red are statistically significant results.

Apart from comparing scores before and after viewing explanations, I also tested, using the 1-tail test, whether the scores for each metric increased for all possible combination of contexts, controlling for whether respondents had viewed the explanations. Mathematically, this test can

be stated as testing for whether $d < 0$ in $d = x - y$, where x and y are the scores from two different contexts. Significant decreases are therefore also accounted for because x and y can be interchanged. For example, I tested whether effectiveness was greater for the PDPC in comparison with the app developer, before respondents viewed the explanations. This yielded $p = 0.166$ (Table 4.3a). I repeated the same tests for scores reported before (Table 4.3) and after viewing the explanations (Table 4.4).

Hence, there are two more hypotheses I test for, setting $\alpha = 0.1$:

H2: There is no increase in scores for [metric A] when [context X] is compared with [context Y].

H3: There is an increase in scores for [metric A] in [context X] is compared with [context Y].

where [metric A] is one of the 4 metrics, and [context X] & [context Y] are all the possible combinations of contexts.

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

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

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

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

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

Increase	Decrease
0.844	0.156

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

TABLE 4.2: p-values comparing mean scores

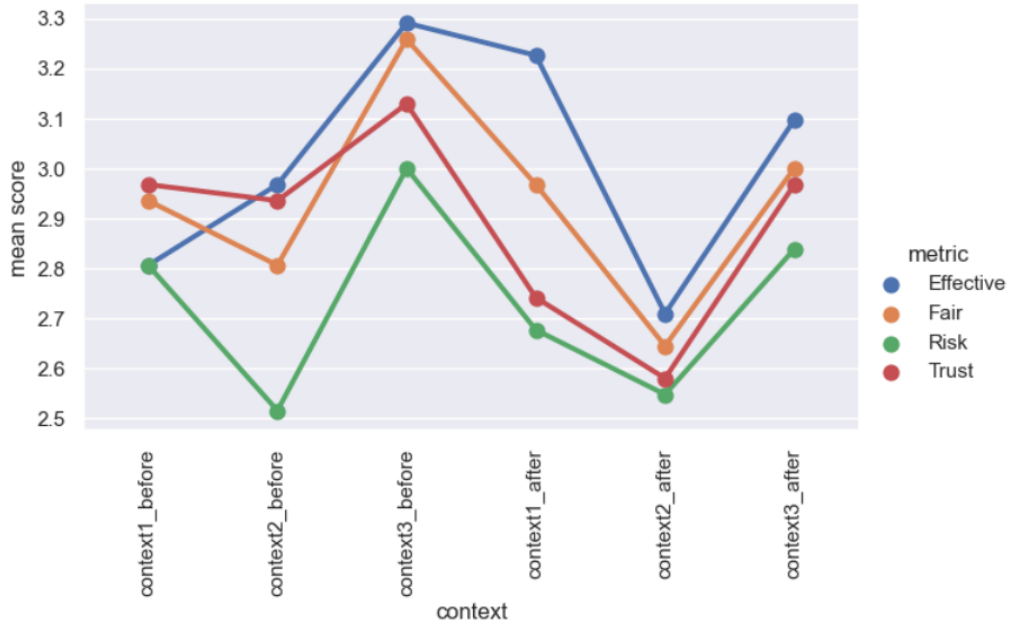


FIGURE 4.3: Mean scores for each metric and each context before and after respondents viewed explanations

Here are the instances when $p < 0.1$, and therefore H_0 and H_2 can be rejected in favour of H_1 and H_3 respectively:

1. Effectiveness and trust significantly decreased for context 2 (PDPC),

Context	App Developer	PDPC	Consumer
App Developer	NA	0.166	0.0128
PDPC	0.834	NA	0.0128
Consumer	0.987	0.987	NA

(A) Effectiveness

Context	App Developer	PDPC	Consumer
App Developer	NA	0.793	0.0909
PDPC	0.207	NA	0.00633
Consumer	0.909	0.994	NA

(B) Fairness

Context	App Developer	PDPC	Consumer
App Developer	NA	0.905	0.236
PDPC	0.0949	NA	0.0188
Consumer	0.764	0.981	NA

(C) Risk

Context	App Developer	PDPC	Consumer
App Developer	NA	0.627	0.194
PDPC	0.373	NA	0.117
Consumer	0.806	0.883	NA

(D) Trust

TABLE 4.3: p-values comparing for *increase* in metric between contexts *before* viewing explanations (i.e. whether scores for context in row < column)

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

2. There is a statistically significant decrease in explainability metrics for context 1 and 2 when taking the mean of the scores across the metrics (Table 4.2a) and for trust across the three contexts (Table 4.2b).
3. Before respondents viewed the explanations, effectiveness (Table 4.3a) and fairness (Table 4.3b) were significantly greater for the consumer

Context	App Developer	PDPC	Consumer
App Developer	NA	0.994	0.771
PDPC	0.00596	NA	0.0192
Consumer	0.229	0.981	NA

(A) Effectiveness

Context	App Developer	PDPC	Consumer
App Developer	NA	0.969	0.443
PDPC	0.0319	NA	0.00248
Consumer	0.557	0.998	NA

(B) Fairness

Context	App Developer	PDPC	Consumer
App Developer	NA	0.822	0.232
PDPC	0.178	NA	0.0673
Consumer	0.769	0.933	NA

(C) Risk

Context	App Developer	PDPC	Consumer
App Developer	NA	0.905	0.108
PDPC	0.0949	NA	0.00845
Consumer	0.892	0.992	NA

(D) Trust

TABLE 4.4: p-values comparing for *increase* in metric between contexts *after* viewing explanations (i.e. whether scores for context in row < column)

in comparison to the app developer, and for the consumer in comparison to the PDPC. Risk (Table 4.3c) was significantly greater¹ for the PDPC in comparison to the consumer, and for the PDPC in comparison to the app developer.

4. After respondents viewed the explanations, effectiveness (Table 4.4a), fairness (Table 4.4b) and trust (Table 4.4d) were significantly greater for the app developer when compared with the PDPC, and the

¹Respondents reported scores from 1 = normatively negative outcome to 5 = normatively positive outcome. For risk, 1 = very risky and 5 = least risky, and an increase in scores when going from context X to Y means that X is riskier than Y.

consumer when compared with the PDPC. Risk was significantly greater for the PDPC when comparing with the consumer (Table 4.4c).

The explanations presented to respondents were deliberately chosen to demonstrate the limitations of the model (see Figure 4.5). Hence, respondents were likely more cognisant of such limitations² when they answered the same questions again in Part 6. Here are some inferences that I draw from these observations, where I mainly focus on the differences in scores before and after viewing explanations in Table 4.1 and 4.2, supported by more specific observations from Table 4.3 and 4.4 when helpful:

1. *"Efficiency" is dependent on the risks of making wrong predictions (app developer, PDPC):* The speed of automation of reading privacy policies would be the same regardless whether an app developer or PDPC uses it. However, changes in efficiency differed between the two contexts (Table 4.1), suggesting that automation was not the only metric that respondents considered part of "effectiveness". Respondents could have considered that automation was more important to app developers when balanced against the risks of making a wrong prediction, as compared to the PDPC, where the risks were too high to justify automation. Hence, the classifier is not "effective" in terms of making a legally defensible decision given that it is fallible. This is also supported by how effectiveness increased for both the app developer and consumer in relation to the PDPC (Table 4.4a).

²This can also be inferred from the negative trend of interpretability and understandability as explained below.

2. *Expertise of the end user of explanations affects explainability metrics:*

In relation to the overall drop in explainability metrics for the contexts of the PDPC and the user (Table 4.2a), perhaps the respondents thought that PDPC and the user did not have technical knowledge to better understand the classifier's logic beyond the explanations that were presented to them, whereas the app developer would have such technical knowledge. Hence, the lack of technical expertise of the PDPC and the user caused the drop in explainability metrics.

3. *Mismatch in expectations particularly affects trust:*

In the case of trust being the only metric that significantly decreased across the contexts (Table 4.2b), this could point to a mismatch of expectations. Intuitively, if respondents do not understand how a process works, they are less likely to trust the results of that process. Perhaps respondents' expectations of AI was that AI would reason similarly to how humans would, or that AI functioned entirely objectively similar to applying a formula in Excel where it is clear how the result was obtained from the formula. Instead, the explanations provided showed that the classifier was inconsistent, sometimes relying heavily on words like "cookies" and other times not so much without any discernible reason. So respondents' expectations were not met, and this caused the overall drop in trust. In fact, viewing the explanations seemed to have decreased trust the most for the PDPC in relation to the app developer and consumer (Table 4.4d), suggesting that trust is a priority for legal institutions like the PDPC.

The intuition behind increasing explainability is to increase trust of users. Trust is especially important to the judicial process as it is closely related to confidence and compliance with court decisions [9]. Hence, this relationship between explainability and trust is worth further investigation because there is evidence that shows this relation is not as general as previously thought [18]. In fact, if explanations reveal problems about the system, trust could decrease rather than increase. This could be a possible explanation of the results because respondents were intentionally given examples that demonstrated the limitations of the classifiers.

4. *All four explainability metrics seem to be most applicable to the PDPC:* Before respondents viewed the explanations, the significant increases of scores when comparing contexts was mostly seen for the consumer in relation to the two other contexts (Table 4.3). However, after viewing the explanations (Table 4.4), the significant increases in scores were only seen for the PDPC in relation to the other two contexts.

A possible interpretation is that respondents, before viewing the explanations, saw these metrics as most relevant to a consumer in relation to the other two contexts. However, after viewing the explanations, respondents saw the metrics as most relevant to the PDPC rather than the other two contexts. Perhaps when comparing between the two types of legal decision-making replaced by the classifier (legal advice of a lawyer for the app developer and consumer contexts, and legal adjudication for the PDPC), respondents

thought that fallible algorithmic decision-making would impact legal adjudication the most across these four metrics.

One strand of analysis not borne by the results is the relation between the consequences of a wrong decision and the level of explainability required. For example, if an algorithmic judge convicted the wrong person for murder, it would have irreversible consequences. Hence, the algorithmic decision-making must be highly transparent. In fact, the AI Act has relied on this relation. It differentiates AI systems into those which are "high risk" and "medium risk", with "high risk" systems requiring a greater level of explainability that is user-empowering and compliance-oriented [28].

However, risk for all the contexts did not significantly change after viewing the explanations except for an increase of risk for the PDPC in relation to the consumer (Table 4.4c). As previously mentioned, I intentionally made sure that the legal consequences of making a wrong decision was capped monetarily at \$10,000 for all contexts because I wanted to reduce the amount of legal knowledge necessary to understand the contexts. However, I did not specifically restrict respondents from assuming that there could be other non-legal consequences. Nevertheless, perhaps the lack of significant result was due to the phrasing of the question that was posed to respondents as "risk to society" which could be interpreted to be a much broader claim rather than just "risk".

Overall, these results confirm that explainability is dependent on the purposes of explanation and values related to explainability can vary depending on the context. This may not be surprising since intuitively, if AI models substitute decision-making previously done by humans, the

models' reasoning process would naturally be compared with the standards of human decision-making. For example, AI that substitutes the decision-making of a secretary to propose mutually convenient meeting times would be expected to prioritise efficiency over transparency. Similarly, algorithmic legal decision-making would have its prioritisation of these values. In particular, as alluded to above, these four explainability metrics seem to be of most importance to algorithmic legal adjudication compared to algorithmic lawyering.

4.2 Part 3: Testing whether viewing more visualisations affected explainability

4.2.1 Analysis of reported interpretability and understandability

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

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

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

	Increase	Decrease
Interpretable	0.867	0.132
Understandable	0.999	<0.001

TABLE 4.5: p-values comparing reported understandability and interpretability between the first and last question

Mean scores of understandability and interpretability after viewing each explanation

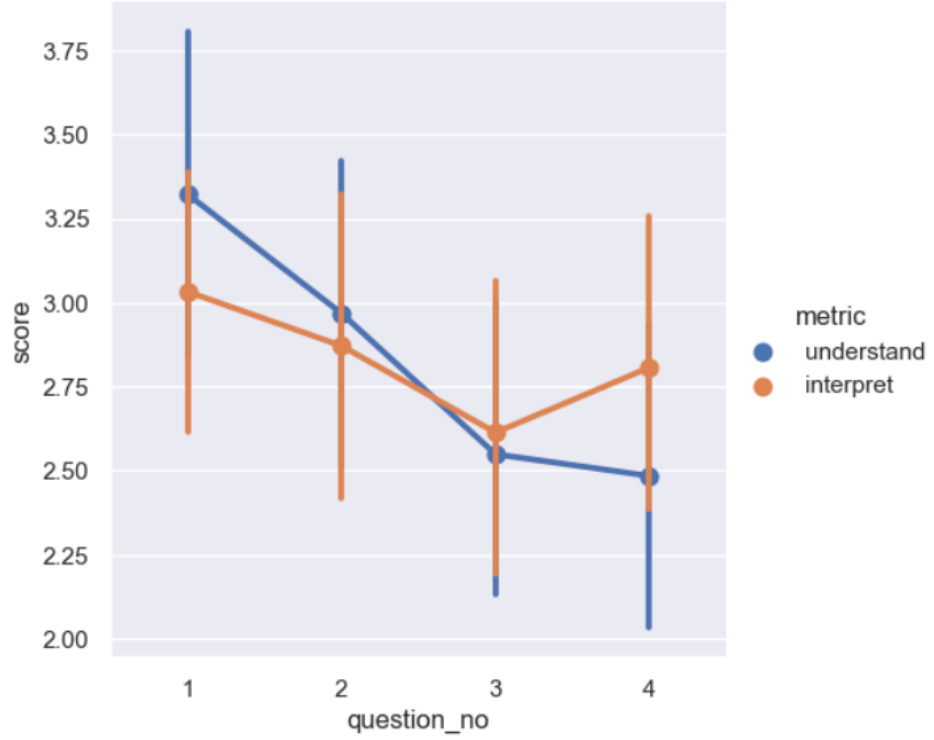


FIGURE 4.4: Trend of the mean of self-reported understanding and interpretability after viewing each explanation

There was a statistically significant decrease in understandability but not interpretability (Table 4.5). Hence, H_0 can be rejected in favour of H_1 .

A reason for why there was no significant decrease in interpretability was because the questions I posed to the respondents were clear in distinguishing the visualisation from the classifier's logic³. Hence, respondents understood how to interpret the visualisation, but were less clear about

³The two questions posed to respondents were: (1) Do you understand why the model made the prediction? and (2) Did you find the visualisation easy to interpret?

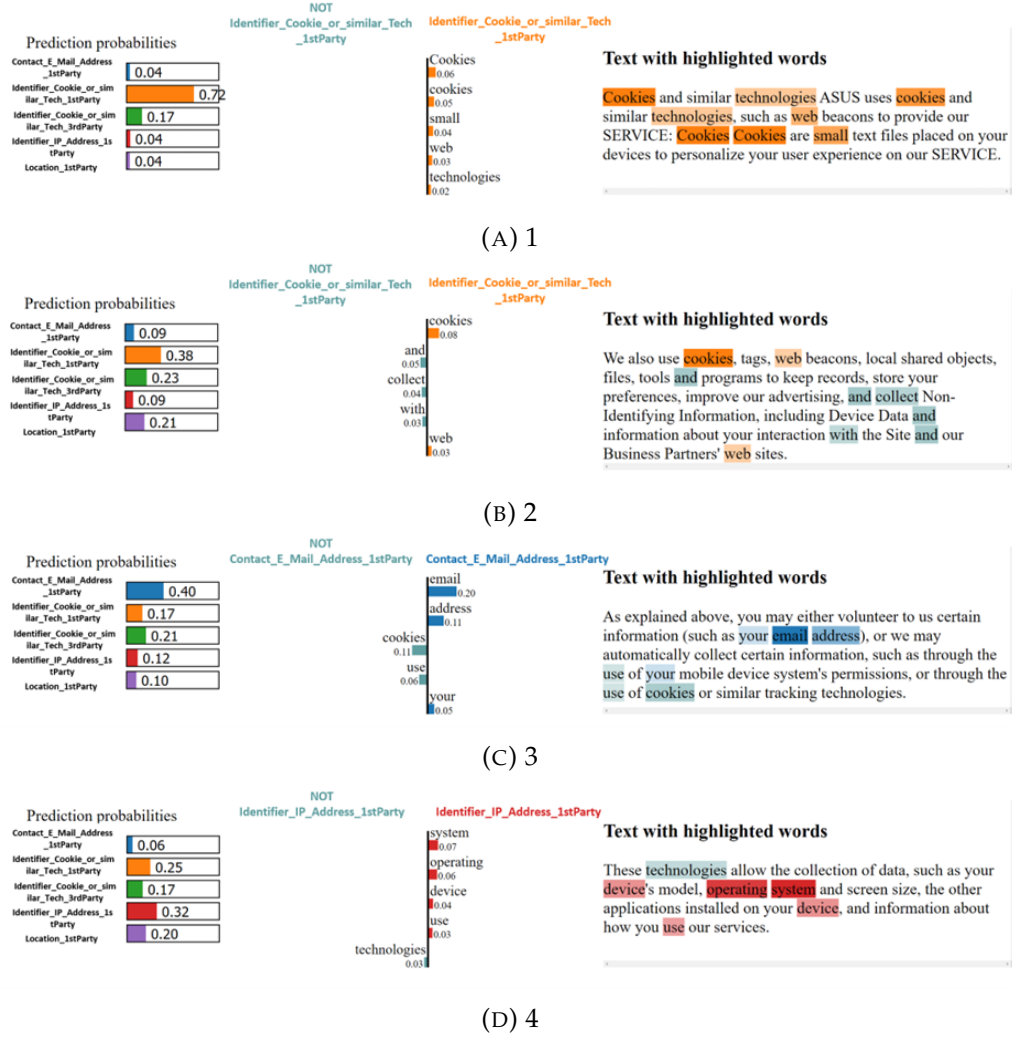


FIGURE 4.5: The 4 visualisations shown to respondents. All 4 sentences were annotated as Identifier Cookie 1st Party. The classifier classified the sentences according to the practice with the highest prediction probability.

the classifier's logic that was presented through the visualisation.

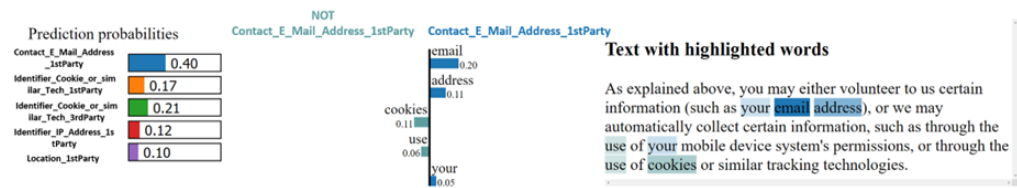
With regards to the significant decrease in understandability, the respondents could be confused about how the model works globally after being exposed to specific predictions that were in themselves understandable, rather than being confused about a specific prediction of the model. As mentioned, the visualisations for this part were specifically chosen to demonstrate the limits of the classifier by changing keywords

which were strong predictors of the data practice. This means that respondents were being exposed to a more complex (and confusing) of the model globally as they found contradictions in how the model used certain keywords in some examples but not in others. Therefore, each explanation was actually effective in communicating to the respondents how that particular prediction was made, and therefore could be considered as locally explainable. However, as a whole, respondents' understandability of the classifier decreased given the contradictions between each explanation. Therefore, the classifier could be said to be locally understandable, but globally less understandable.

Another possible inference is that the underlying explainability method also influences the respondents' perception of the interpretability of the visualisation technique (or vice versa). This is seen through the positive correlation with a negative gradient between the understand and interpret scores. This is not surprising since if the visualisation technique is unclear or poor, understandability of the classifier itself would also decrease since respondents' only way of viewing the results of the explainability technique is through the visualisation technique.

4.2.2 Analysis of predicting counterfactuals

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



(A) Original explanation



(B) Counterfactual explanation (not shown to respondents)

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

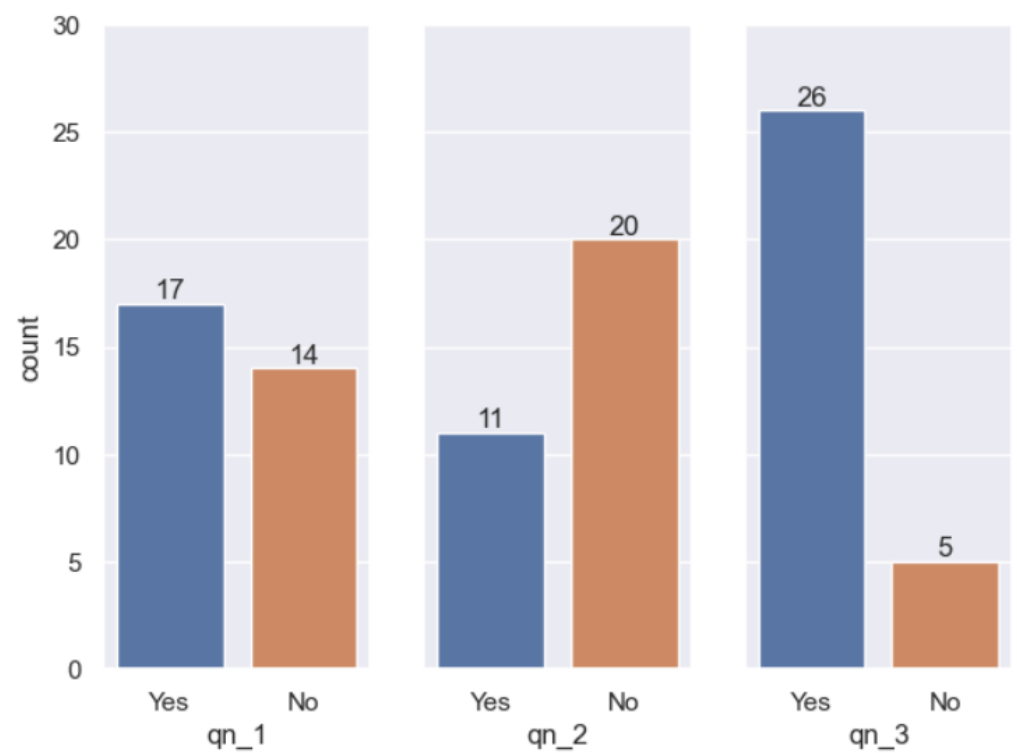


FIGURE 4.7: Votes for predicting whether counterfactual would be classified as Identifier_Cookie_1st_Party

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

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

Overall, respondents found no difference in understandability between logistic regression and SVC (Figure 4.8), while it was more contentious when comparing text representations, with a third split across the three options (Figure 4.9).

Remember that for specific **class performance** of Identifier Cookie 1st Party, the performance difference between the classifiers that controlled for SVC but varied the text representation was higher than the performance difference for classifiers that controlled for GloVe but varied the model. Assuming that higher classifier performance correlates with higher understandability, when comparing respondents' votes with the "ground truth" metric of classifier performance, most respondents would be expected to indicate that: (1) there were differences in the understandability when comparing SVC + Tf-IDF vs SVC + GloVe, and (2) no differences for SVC + GloVe vs logistic regression + GloVe.

This second expectation is validated (Figure 4.8) since the overwhelming majority of respondents chose "no difference". For the first expectation, while there were no majority votes for one option, votes were almost equally split instead of being heavily weighted in favour of one option (Figure 4.9). This could suggest that respondents were more undecided about the understandability of the classifiers which somewhat

reflects their similar performance figures. Overall, these results provide some support for some relation between understandability and classifier performance.

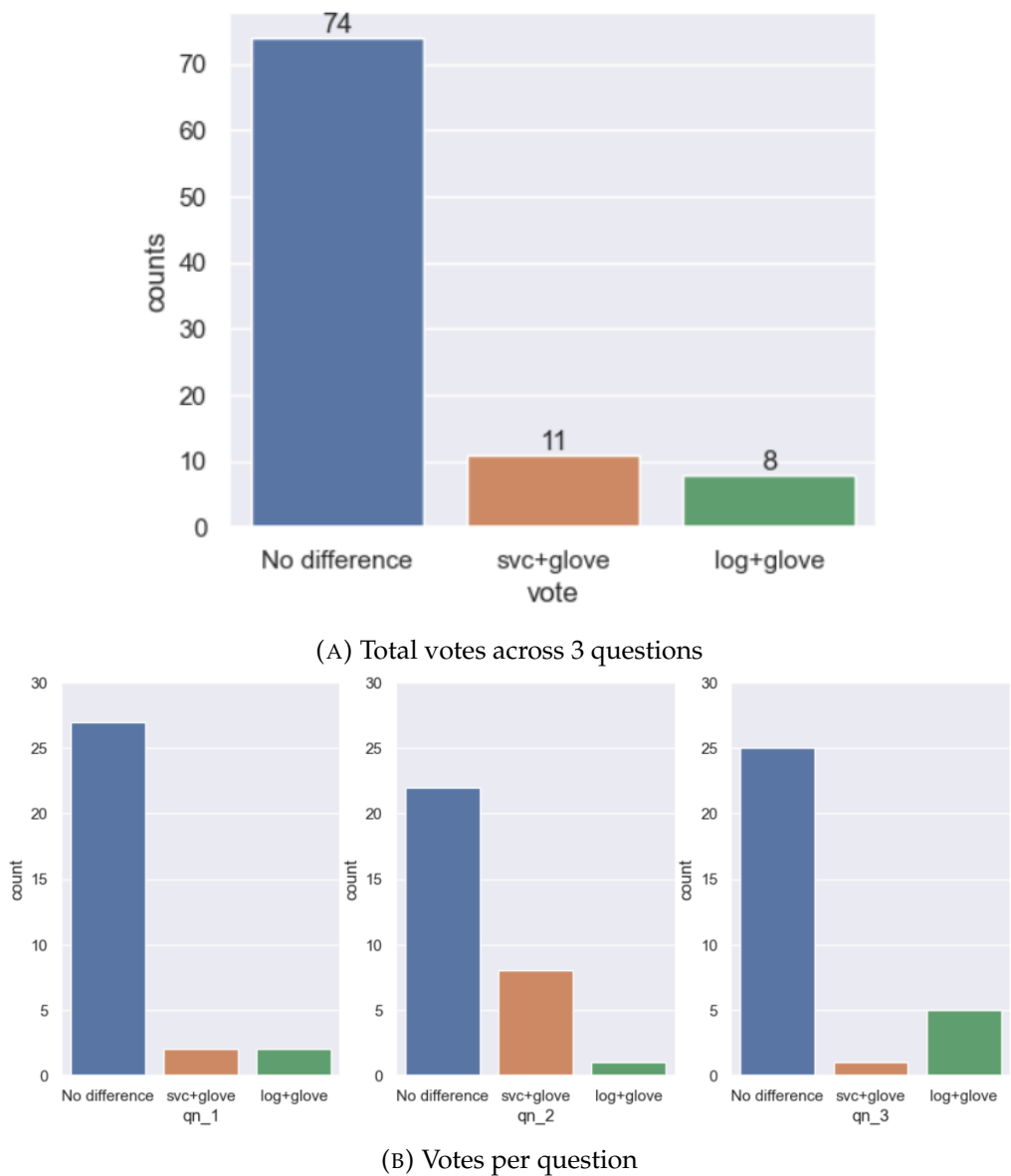


FIGURE 4.8: Respondents’ votes to whether SVC + GloVe (55% F1) or Logistic regression + GloVe (59% F1) were more understandable. F1 scores refer to specific class performance for Identifier Cookie 1st Party.

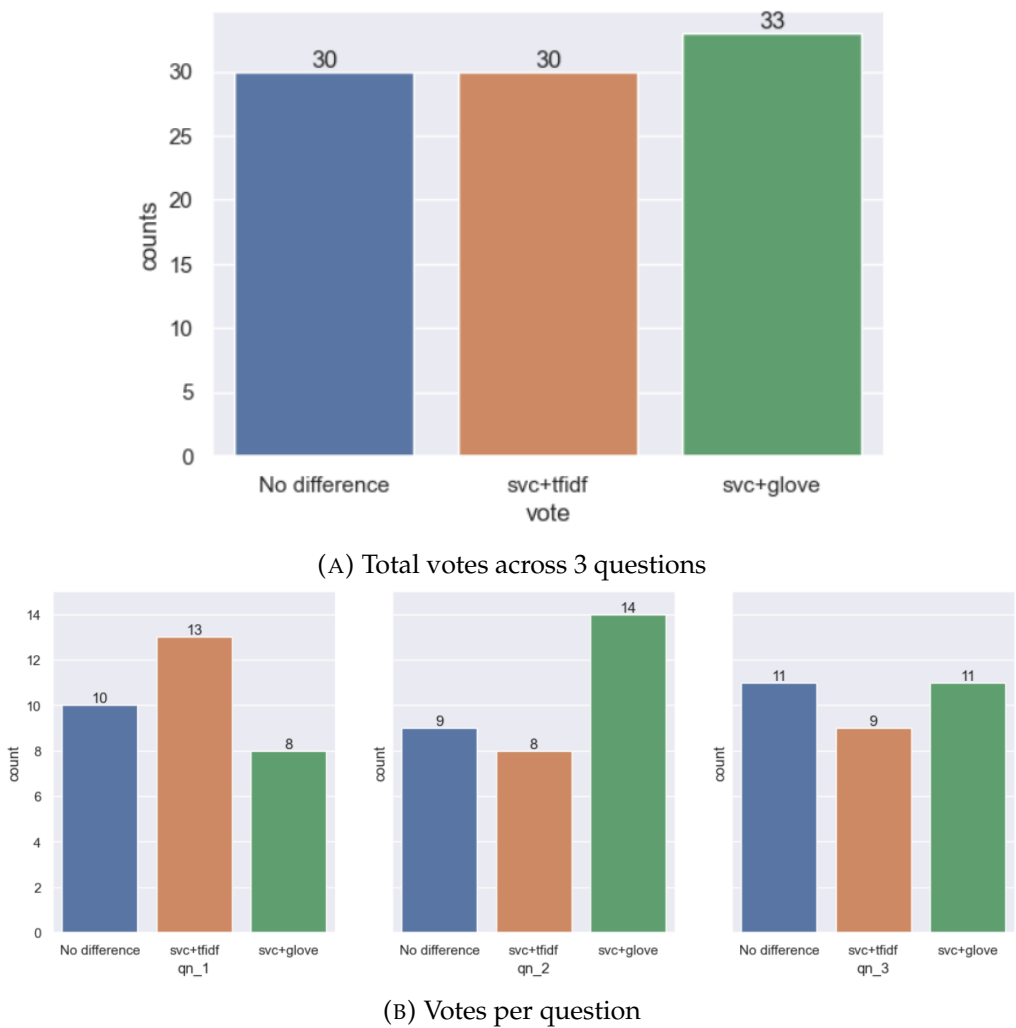


FIGURE 4.9: Respondents' votes to whether SVC + Tf-IDF (68% F1) or SVC + GloVe (55% F1) were more understandable. F1 scores refer to specific class performance for Identifier Cookie 1st Party.

5 Conclusion

5.1 Findings

The **research questions** can now be answered:

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

Across all classes, SVC + Tf-IDF performed the best when measuring performance using weighted F1, though there was not much difference when comparing micro-averaged AUC for the ROC curves. In terms of performance specifically for Identifier Cookie 1st Party, both SVC + Tf-IDF and logistic regression + Tf-IDF performed similarly. It is surprising that GloVe embeddings performed worse overall and caused greater variance in the AUC of all the classes compared to Tf-IDF, given that GloVe accounts for semantic context as compared to Tf-IDF which is purely count-based.

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

When SVC was compared with logistic regression, respondents by a large majority found no difference in their understandability. When Tf-IDF was compared to GloVe, respondents were undecided

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

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

Understandability of why the classifier made such predictions significantly decreased. However, interpretability of the visualisation produced by LIME did not significantly decrease. Both understandability and interpretability were positively correlated with a negative gradient.

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 and the PDPC?

Effectiveness and trust significantly decreased for the PDPC context, and fairness for the consumer context. Only effectiveness significantly increased for the app developer context. Explainability metrics in general significantly decreased specifically for the PDPC and consumer context. When taking the mean scores across the contexts, trust significantly decreased.

Before respondents viewed the explanations, effectiveness and fairness were significantly greater for the consumer in comparison to the app developer, and for the consumer in comparison to the PDPC. Risk was significantly greater for the PDPC in comparison to the consumer, and for the PDPC in comparison to the app developer.

After respondents viewed the explanations, effectiveness, fairness and trust were significantly greater for the app developer when compared with the PDPC, and the consumer when compared with the PDPC. Risk was significantly greater when comparing the PDPC with the consumer.

As mentioned in the **survey design**, this capstone also serves as a trial for XAI human evaluation that assesses explainability across different selected purposes of explanations. Overall, I would consider this survey design a success, given the statistically significant results reported.

5.2 Limitations and future work

There are a few limitations:

1. *Limited number of respondents ($n = 31$), diversity of expertise and age:*

The original intent was to conduct a more in-depth cross-sectional study of the self-reported scores of the respondents according to their area of expertise and their beliefs relating to AI and data privacy. This analysis could be used to support certain inferences made earlier such as explainability is dependent on expertise of the end user. However, with 31 respondents, breaking down the survey

results into smaller sub-groups would not reliably show statistical trends¹. Hence, future work could be conducted on a much larger sample size to run cross-sectional analysis.

2. *Analysis is limited to the context of the APP-350 corpus, classifiers used, and LIME:* Assessing XAI is highly context dependent, and therefore the foregoing analysis should also be seen in light of these particular components. Future work to evaluate XAI within the context of data privacy could include using another data privacy dataset, other classifiers and other XAI techniques apart from LIME. In particular, LIME may interact differently with current state-of-the-art classifiers which are much more complex and rely on deep learning given that NLP has developed quite substantially since 2016 when LIME was first introduced.
3. *Inherent drawbacks of LIME and human evaluation of XAI:* As LIME is a post-hoc local explainability technique, respondents' understanding of the model was limited to the example visualisations that were presented to them. The sentences shown to respondents were intended to show the limitations of the classifier. This paints a more confusing and inconsistent picture of the classifier's logic, as compared to examples that consistently show "cookie" was the most important feature. In fact, this was implemented in another study where the authors only chose to show correctly classified sentences in order to reduce the information load on the respondents [10].

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

If local techniques are used, future work could include an interactive dashboard where respondents can choose to generate their own sentences for classification, which allows them to test their understanding of the classifier's logic.

Further, human evaluations of post-hoc explanations have been criticised for being inherently flawed because of confirmation bias and therefore such evaluations may have limited meaning [26]. Instead, [26] 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.

Bibliography

- [1] Hussam Alkaissi and Samy I McFarlane. “Artificial hallucinations in chatgpt: Implications in scientific writing”. In: *Cureus* 15.2 (2023).
- [2] Kevin D. Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017.
DOI: [10.1017/9781316761380](https://doi.org/10.1017/9781316761380).
- [3] Patricia Bassey. *Logistic Regression vs Support Vector Machines (SVM)*.
Last Accessed: 2023-03-21. 2019. URL: <https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16>.
- [4] Simon Chesterman. “We, The Robots: Regulating Artificial Intelligence and the Limits of the Law”. In: Cambridge University Press, 2021. Chap. 6. Transparency.
- [5] Simon Chesterman. “We, The Robots: Regulating Artificial Intelligence and the Limits of the Law”. In: Cambridge University Press, 2021. Chap. 3. Opacity.
- [6] Marina Danilevsky et al. “A Survey of the State of Explainable AI for Natural Language Processing”. In: *CoRR* abs/2010.00711 (2020).
arXiv: [2010.00711](https://arxiv.org/abs/2010.00711). URL: <https://arxiv.org/abs/2010.00711>.

- [7] Finale Doshi-Velez and Been Kim. “Towards a Rigorous Science of Interpretable Machine Learning”. In: *arXiv preprint arXiv:1702.08608* (2017).
- [8] Julia El Zini and Mariette Awad. “On the Explainability of Natural Language Processing Deep Models”. In: *ACM Computing Surveys* 55.103 (2022), pp. 1–31. DOI: <https://doi.org/10.1145/3529755>.
- [9] Jane Goodman-Delahunty. “Measuring Trust and Confidence in Courts”. In: (2021).
- [10] Łukasz Górski and Shashishekar Ramakrishna. “Explainable Artificial Intelligence, Lawyer’s Perspective”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL ’21. São Paulo, Brazil: Association for Computing Machinery, 2021, 60–68. ISBN: 9781450385268. DOI: [10.1145/3462757.3466145](https://doi.org/10.1145/3462757.3466145). URL: <https://doi.org/10.1145/3462757.3466145>.
- [11] Stanley Greenstein. “Preserving the rule of law in the era of artificial intelligence (AI)”. In: *Artificial Intelligence and Law* 30.3 (2022), pp. 291–323.
- [12] Oskar J. Gstrein and Anne Beaulieu. “How to protect privacy in a datafied society? A presentation of multiple legal and conceptual approaches”. In: *Philos Technol* 35.1 (2022), p. 3. DOI: <https://doi.org/10.1007%2Fs13347-022-00497-4>.
- [13] Balint Gyevnar and Nick Ferguson. “Aligning Explainable AI and the Law: The European Perspective”. In: *arXiv preprint arXiv:2302.10766* (2023).

-
- [14] Philipp Hacker and Jan-Hendrik Passoth. “Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond”. In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer. 2022, pp. 343–373.
 - [15] Bi-Min Hsu. “Comparison of supervised classification models on textual data”. In: *Mathematics* 8.5 (2020), p. 851.
 - [16] IBM. *What is natural language processing?* Last Accessed: 2023-02-11. 2022. URL: <https://www.ibm.com/sg-en/topics/natural-language-processing>.
 - [17] Javatpoint. *Logistic Regression in Machine Learning*. Last Accessed: 2023-03-21. n.d. URL: <https://www.javatpoint.com/logistic-regression-in-machine-learning>.
 - [18] Lena Kästner et al. “On the relation of trust and explainability: Why to engineer for trustworthiness”. In: *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE. 2021, pp. 169–175.
 - [19] Daniel Martin Katz et al. “GPT-4 Passes the Bar Exam”. In: *Available at SSRN* 4389233 (2023).
 - [20] Zhiyuan Liu et al. “Word representation”. In: *Representation Learning for Natural Language Processing* (2020), pp. 13–41.
 - [21] Alessandro Mantelero. “The future of consumer data protection in the EU Re-thinking the “notice and consent” paradigm in the new era of predictive analytics”. In: *Computer Law & Security Review* 30.6 (2014), pp. 643–660.

-
- [22] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. Last Accessed: 2023-02-11. 2022. URL: <https://openai.com/blog/chatgpt/>.
- [23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: <http://www.aclweb.org/anthology/D14-1162>.
- [24] Personal Data Protection Commission. *Model Artificial Intelligence Governance Framework Second Edition*. Tech. rep. Last Accessed: 2023-2-20. Personal Data Protection Commission, Jan. 2020. URL: <https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf>.
- [25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. DOI: 10.48550/ARXIV.1602.04938. URL: <https://arxiv.org/abs/1602.04938>.
- [26] Avi Rosenfeld. "Better Metrics for Evaluating Explainable Artificial Intelligence". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '21. Virtual Event, United Kingdom: International Foundation for Autonomous Agents and Multiagent Systems, 2021, 45–50. ISBN: 9781450383073.
- [27] SciPy. *scipy.stats.wilcoxon*. Last Accessed: 2023-3-1. 2023. URL: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html>.

-
- [28] Francesco Sovrano et al. “Metrics, explainability and the European AI act proposal”. In: *J* 5.1 (2022), pp. 126–138.
- [29] Daniel Vale, Ali El-Sharif, and Muhammed Ali. “Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law”. In: *AI and Ethics* (2022), pp. 1–12.
- [30] Giulia Vilone and Luca Longo. “Notions of explainability and evaluation approaches for explainable artificial intelligence”. In: *Information Fusion* 76 (2021), pp. 89–106.
- [31] Ekaterina Vylomova et al. *Take and Took, Gaggles and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning*. 2015. DOI: [10 . 48550 / ARXIV . 1509 . 01692](https://doi.org/10.48550/ARXIV.1509.01692). URL: [https :
//arxiv.org/abs/1509.01692](https://arxiv.org/abs/1509.01692).
- [32] Isabel Wagner. “Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021”. In: *arXiv preprint arXiv:2201.08739* (2022).
- [33] Carlos Zednik. “Solving the black box problem: A normative framework for explainable artificial intelligence”. In: *Philosophy & technology* 34.2 (2021), pp. 265–288.
- [34] Sebastian Zimmeck et al. “MAPS: Scaling Privacy Compliance Analysis to a Million Apps”. In: *Proceedings on Privacy Enhancing Technologies* 2019.3 (2019), pp. 66–86.