Automatic Classification of Legal Violations in Cookie Banner Texts

Marieke van Hofslot

A thesis presented for the MSc of Artificial Intelligence

Utrecht University

Almila Akdag Salah
Albert Gatt

December 2022

Abstract

Cookie banners are designed to request consent from website visitors for their personal data. Recent research suggest that a high percentage of cookie banners violate legal regulations as defined by the General Data Protection Regulation (GDPR) and the ePrivacy Directive. In this paper, we focus on language used in these cookie banners, and whether these legal violations can be automatically detected. We make use of a small cookie banner dataset that is annotated by five experts for legal violations and test it with state-of-the-art classification models, namely BERT, LEGAL-BERT, BART in a zero-shot setting, and BERT with LIWC embeddings. Our results show that none of the models outperform the others in all classes, but in general, BERT and LEGAL-BERT provide the highest accuracy results (70%-97%). However, even these best performing models are influenced by the the unbalanced distributions in the dataset.

Contents

1 Introduction **4**
1.1 Research question 6

2 Dark Patterns **7**
2.0.1 Dark pattern classification 7
2.0.2 Dark patterns in cookie banners 9
2.0.3 GDPR 9
2.0.4 Cookie banners compliance with the GDPR 10

3 Natural Language Processing **12**
3.1 Natural Language Processing Architectures 12
3.1.1 Basics of NLP 12
3.1.2 Machine Learning 14
3.1.3 Neural networks 15
3.1.4 BERT 18
3.1.5 Fine-tuned BERT 18
3.1.6 Conclusion 19
3.2 Zero shot learning 19
3.2.1 Conclusion 20
3.3 Stylistic classification 20
3.3.1 Authorship detection 20
3.3.2 Sentiment analysis 21
3.3.3 Deception detection 21
3.3.4 LIWC 21
3.3.5 Conclusion 22

4 Data **23**
4.1 Dataset 23
4.2 Annotation classes and classification labels 23
4.3 Legal violation and annotation 24

4.4 Challenges data 25

5 Method **27**
5.1 Data preprocessing 27
5.1.1 BERT with LIWC features 29
5.1.2 BART in ZS-setting 30
5.2 Models 32
5.2.1 BERT 33
5.2.2 BERT with LIWC features 33
5.2.3 LEGAL-BERT 33
5.2.4 BART in ZS-setting 33
5.2.5 Training details and hyperparameters 34
5.3 Evaluation 34

6 Results **35**

7 Discussion **40**

8 Conclusion **42**

9 Appendix **43**
9.1 Ethical implications and limitations 43
9.2 Full results 44

Chapter 1

Introduction

Dark patterns are user interface design choices that coerce users into making unintended decisions. One example of dark patterns can be found in cookie banners - banners that prompt users to consent to the use of cookies as described in their cookie policy. Website operators are required to be transparent and explain the purpose of cookie use, but recent research shows that 89% of cookie banners violate applicable laws (Santos et al., 2021). There is little research about automatic detection of dark patterns in cookie banners and so far, the legality of cookie banners has been manually annotated. The aim of this project is thus to use natural language processing to create a method of automatic detection of dark patterns in cookie banners according to their legal violations.

Since the term was first introduced in 2015 (Brignull et al., 2015), dark patterns have received a lot of attention within various research communities. Although there is variation in classification (Gray et al., 2018) and definition of dark patterns (Mathur et al., 2021), one of the main facets of dark patterns is to have user interface properties which can affect a user. Furthermore, another important part of the definition is that the designer intentionally deploys dark patterns to accomplish a certain goal. Sometimes dark patterns aim simply to benefit an online service (Utz et al., 2019), but commonly it involves harm to users (Gray et al., 2018), such as causing financial loss or tricking users into giving away their personal data (Stavrakakis et al., 2021). This goal to harm is also reflected in the differentiation between dark patterns and anti-patterns; whereas anti-patterns arise from lack of skill from the designer, in dark patterns there is malice assumed. (Mathur et al., 2021). Dark patterns are especially concerning since they are very effective at getting people to make choices they do not intend to make (Luguri & Strahilevitz, 2021; Narayanan et al., 2020) and because users may fail to notice that any nudging mechanism or decoy is present (Mathur et al., 2021).

Although they are gaining an increasing amount of attention in research and they are a common occurrence on the internet, detection remains a challenge due to a large variation in types and implementation of dark patterns (Curley et al., 2021). Stavrakakis et al. (2021) has conducted an examination of possible manual and automated ways of detecting dark patterns. Dark patterns that are easier detectable are patterns that include certain phrases or images that are easy to identify (automatically), such as *trick questions* ("opt in", "opt out" or pre-ticked checkboxes) or *roach motel* ("activate" or "subscribe" links/buttons but no "deactivate" or "unsubscribe"). Patterns that are harder to detect but can be identified manually include *sneak into basket* (sneakily adding additional items to basket) and *hidden cost* (unexpected changes in charges). 'Undetectable' dark patterns, such as *misdirection* (distract attention), *confirmshaming* (guilting user into opting in), *bait and switch* (undesirable thing happens instead of intended thing), or *privacy suckering* (tricked into sharing more information than intented), have too much variation in their definition or implementation, which makes it virtually impossible to detect using web crawling and web scraping techniques (Stavrakakis et al., 2021).

Dark patterns also appear in cookie banners. These banners appear when a user visits a website for the first time, and requests authorisation to the use of cookies and other trackers for a range of purposes. To comply with the General Data Protection Regulation (GDPR) (EU, 2018) and the ePrivacy Directive (EU, 2009), website operators have to inform EU users and ask for their consent for the processing of their personal data for 'unnecessary purposes', i.e. data that is not needed for the website to function, such as user-targeted advertising (Article 29 Working Party, 2012). A consent request needs to be unambiguous, clear, concise, and informative, and consent needs to be freely given (Articles 4(11) and 7(2,) (EU, 2018)). Despite the legal requirements described in the GDPR, research shows that there are still a lot of legal violations and dark patterns in cookie banners (Santos et al., 2021; Soe et al., 2020), privacy policies (Degeling et al., 2018) and consent management platforms (Nouwens et al., 2020)). The legal study by Santos et al. (2021) focused on processing purposes of cookie banners and confirmed that 89% of the cookie banners violated at least one legal requirement applied to the text of the stated purposes; they further detected the use of vagueness, framing, misleading wording, and technical jargon. Utz et al. (2019) noted that the text to explain the purpose of data collection was typically expressed in generic terms, and use of technical jargon was not properly understandable by the average data subject. Studies furthermore confirmed that the prevalence of "affirmative" options and positive framing could nudge users toward consenting to tracking (Hausner & Gertz, 2021; Kampanos & Shahandashti, 2021).

There is a need to identify such textual violations and develop tools that can automatically detect such textual *dark patterns* (Mathur et al., 2019b) in order to provide proof of such practices (and legal evidence) to support the legal proceedings of en-

forcement authorities in their auditing efforts. Regulators are presently overwhelmed by the novelty and sheer scale at which such patterns are being deployed online. However, only a few studies have investigated automatic detection of legal violations in cookie banner text. Bollinger et al. (2022) used feature extraction and ensembles of decision trees for their cookie purpose classifier with which they developed a browser extension to remove cookies according to user preferences. Khandelwal et al. (2022) used a fine-tuned BERT Base-Cased model to discover and force cookie settings to disable all non-essential cookies. These studies focus on enhancing the usability of websites for the users, whereas our aim is to detect the legal violations for the purpose of supporting legal proceedings of authorities for auditing.

1.1 Research question

In our research, we focus on automatic detection of legal violations in cookie banner texts. In Santos et al. (2021), cookie banner texts were manually annotated on six legal requirements and the corresponding violations from the GDPR. Their annotation is very thorough, but was very time-consuming and required legal expertise. Our project uses their annotations to investigate whether the detection dark patterns in cookie banners texts can be automated using current state of the art NLP models. Although there might be specific wordings and phrases that can be easily automatically detected (similar to some dark patterns in Stavrakakis et al. (2021)), it might be challenging due to the nature of the text, which is meant to confuse users. Our aim is to understand if language models can be used with little or no fine-tuning for auditing purposes by policymakers or consumer protection organisations.

We first look into which models are available and appropriate to use for the automatic detection of legal violations of cookie banner texts. Our first research question is formulated as follows:

- **RQ1:** Which NLP models are suitable to use for the automatic detection of legal violations in cookie banners?

Furthermore, we look into the performance of the chosen models and evaluate their performance according to several evaluation metrics. The second research question is thus formulated as:

- **RQ2:** How well do the models perform in terms of classification accuracy?

Ultimately, we are interested in whether the models produce reliable enough accuracy scores to use for auditing purposes. In addition to reporting the models' performance, we will document the strengths and shortcoming of the models to provide insight on the challenges of such a classification task.

Chapter 2

Dark Patterns

The term 'dark patterns' was first coined by Harry Brignull and was then defined as "tricks used in websites and apps that make you do things that you didn't mean to, like buying or signing up for something" (Brignull et al., 2015). Since then, there has been a lot of attention to dark patterns. The literature, however, has some inconsistencies and contradictions in the definition and types of dark patterns. Mathur et al. (2021) reviewed literature about dark patterns in Human-Computer Interaction (HCI) and found that there is significant variation in how dark patterns are defined. They came up with four facets of dark pattern definitions that are used in the literature to define dark patterns. First, dark patterns have user interface properties that can affect users, which can be misleading, coercing or deceiving for users. Secondly, dark patterns have a mechanism of effect for influencing users. Some definitions describe this as subverting user preferences (Bösch et al., 2016; Mathur et al., 2019a). Furthermore, the third facet of dark patterns is the user interface designer: several definitions state that "the designers intentionally deploy dark patterns to achieve a goal" (Mathur et al., 2021). Lastly, the fourth fact describes the benefits and harms that results from the user interface design. For instance, dark patterns are defined as aiming to benefit an online service (Gray et al., 2020; Utz et al., 2019) or involving harm to users (Gray et al., 2018).

2.0.1 Dark pattern classification

Prior work on dark patterns has come up with various classifications of user interface types with dark patterns. Brignull et al. (2015) originally introduced 12 types of dark patterns: trick question, sneak into basket, roach motel, privacy zuckering, price comparison prevention, misdirection, hidden costs, bait and switch, confirmshaming, disguised ads, forced continuity and friend spam. Misdirection, for instance, is a de-

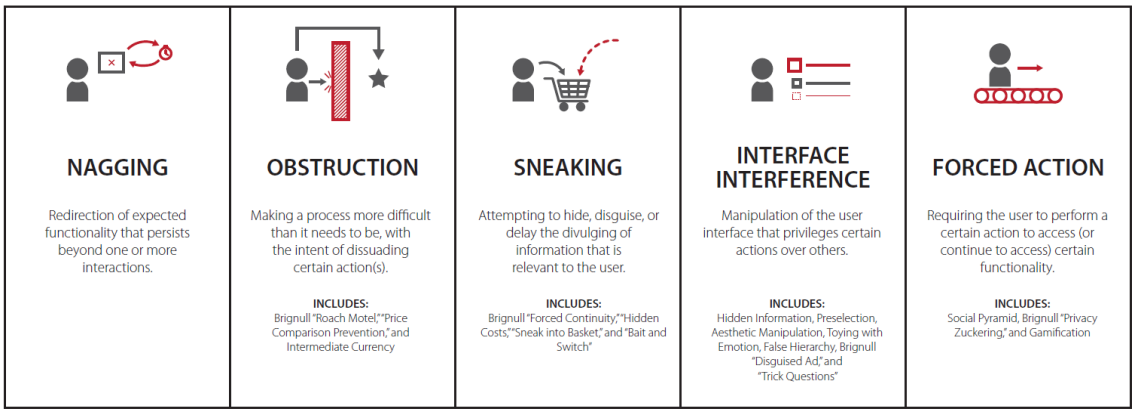


Figure 2.1: Summary of dark pattern strategies from Gray (2018)

sign that purposefully focuses the user's attention on one thing in order to distract their attention from another, and confirmshaming is used to guilt the user into opting into something by using wording in such a way that the user is shamed into compliance. Bösch et al. (2016) introduced seven types of dark patterns specifically on privacy, based on Hoepman's privacy design strategies: maximise, publish, centralise, preserve, obscure, deny, violate and fake. According to Bösch, obscure is used to make it hard or impossible for data subjects to learn how their personal data is collected and processed, which is similar to 'privacy zuckering': named after Facebook CEO Mark Zuckerberg, this dark pattern tricks a user into publicly sharing more information about themselves than the user intended to. Furthermore, Gray et al. (2018) derived a new taxonomy of dark patterns by analysing a corpus of dark patterns collected from users on Twitter. This taxonomy consists of nagging, obstruction, sneaking, interface interference and forced action (see Figure 2.1). The obstruction dark pattern strategy, for example, is defined as "making a process more difficult than it needs to be, with the intent of dissuading certain action(s)" (Gray et al., 2018) and includes Brignull's roach motel and price comparison prevention.

Why use dark patterns?

Although dark patterns have recently gotten more attention, they are the result of other trends that have been used for a long time, such as *nudging*, *growth hacking* and *A/B testing*. Nudging can be described as a way to 'nudge' people in a certain direction by using tactics that play into our psychological biases. Although nudging first arose from research to understand irrational decisions (Tversky & Kahneman, 1974), nowadays businesses use nudging to interact with customers. Growth hacking is used to rapidly increase the growth of a company. Creative design, marketing

and various techniques on data and automation are used to drive product adoption. Another 'weapon' used in marketing is A/B testing. By letting users interact with two or more randomly selected variants of websites, designer found out that even small or trivial changes in design can lead to to differences in user's behaviour. This "idea of data-driven optimatization" (Narayanan et al., 2020) of interfaces has become deeply rooted into the design process of online services, since A/B testing is a massively useful tool to experiment with design choices and their possible influence on customer's behaviour. The use of A/B testing is not necessarily a bad thing. They are, however, becoming a key to the evolution of dark patterns, because it might be useful to help a business attracts more customers, but with certain design choices they make use of dark patterns to mislead and deceive customers into decisions that they did not intend to make.
The one thing that nudging, growth hacking, A/B testing and dark patterns have in common is that they work — at least at the short term (Brownlee, 2016). For example, a towel reuse message in hotels ("75% of guests in this hotel usually use their towels more than once") is effective because it makes use of social norms to get people to alter their actions (Narayanan et al., 2020). Most shopping sites use various kinds of design choices in order to persuade customers into buying, including dark patterns: out of 11K shopping sites, around 11% used dark patterns (Mathur et al., 2019a). Results from Luguri and Strahilevitz (2021) show that dark patterns are effective: users exposed to dark patterns were far more likely to subscribe to a dubious service than users in the control groups, especially when they were exposed to 'aggressive' dark patterns.

2.0.2 Dark patterns in cookie banners

Apart from shopping sites (Mathur et al., 2019a), mobile applications (Di Geronimo et al., 2020) and games (Goodstein, 2021; Zagal et al., 2013), dark patterns can also be found in cookie banners. Cookie banners appear when a user visits a website for the first time and requests authorisation to use cookies and other trackers for a range of purposes.

2.0.3 GDPR

These banners appear due to the European Union's General Data Protection Regulation (GDPR) that went into effect in May 2018. According to the GDPR, users in the EU must be informed about the gathering of their personal data by website operators. Only when cookies and similar tracking technologies are used for non-essential purposes, such as advertising, user consent is required. Even though cookies

are required for the website to function (i.e., necessary functions), website operators must be transparent and explicitly explain the purpose of their use of cookies. Users cannot consent to the collection of their personal data unless they are aware of the explicit purpose(s) of the use of cookies. Article 6 of the GDPR includes that "the data subject has given consent to the processing of his or her personal data for one or more specific purposes" (EU, 2018). Furthermore, Recital 32, Conditions for Consent, states that consent should be a "freely given, specific, informed and unambiguous indication of the data subject's agreement" and that "the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided" (EU, 2018).

2.0.4 Cookie banners compliance with the GDPR

The implementation of the GDPR in 2018 has brought attention to the compliance of cookie banners, cookie consent notices and privacy policies with the GDPR and possible problematic patterns. The aim of Soe et al. (2020) was to study the extent of dark patterns usage specifically concerning the designs used to elicit informed consent. They analysed a manually collected dataset of 300 cookie consent notices from online news services and specifically looked at the existence and variety of dark patterns, possibility for user to not give consent, the location of consent notice on the screen and the complexity of the consent notice. They found a variety of dark patterns that try to find a way around the GDPR by design. Obstruction and interface interference, following the categorisation from Gray et al. (2018), were the dominant pattern types in the collected consent notices. Moreover, the results show that the industry does not have a standard terminology when it comes to the cookie consent notices, which makes it hard for a user to understand what they are giving consent to. There is also a trend to avoid using a 'negative' word for denial of consent: instead, terms like "Read more", "More information" or "Cookie policy" are used instead of a direct wording like "Reject" or "Opt out".

The GDPR states that website operators should be transparent about the purposes of cookie use; only 125 sites out of 300 listed the purpose of the use of cookies. Utz et al. (2019) ran several experiments to determine whether the design of a cookie consent notice influenced the users' consent decision. They specifically looked at the position, number of choices and emphasis/selection, and presence of a privacy policy link or (non-)technical language. Their results provides evidence for dark patterns interfering with users' decisions, since nudging (such as by pre-selecting check boxes) considerably influences users' acceptance of cookies. Concerningly, almost 25% of participants believed that they had to accept cookies before they could gain access to a website. In addition, cookie banners currently have such an abundance in combinations of

information provision, the enforcement of users' choices and user options that there seems to be no improvement for user privacy when comparing to the time before the GDPR went into effect.

Nouwens et al. (2020) also found that dark patterns are still found after introducing the GDPR. Only 11.8% of the five most popular consent management platforms (CMPs) had the minimal requirements stated by the GDPR. The majority of CMPs made it more difficult to reject all tracking than to accept it, a clear example of the obstruction dark pattern. Even if pre-checked boxes are specifically prohibited by the GDPR, a substantial amount of CMPs had pre-ticked boxes of various types. Research by Degeling et al. (2018) shows similar results regarding the compliance with the GDPR of cookie consent. After analysing 6,579 privacy policies in 24 different languages, the authors conclude that although websites became more transparent after the GDPR went into effect, the use of tracking and cookies appeared to be predominantly unchanged. The GDPR's most notable effect that was observed is the increase of cookie consent notifications: from a percentage of 46.1% in January 2018 to 62.1% in May 2018. However, a majority of the analysed cookie consent libraries did not meet GDPR requirements.

Santos et al. (2021) presented an in-depth analysis of cookie banners and their legal compliance with the GDPR with a focus on the purposes of cookie banners. They classified six legal requirements applicable to cookie banner texts: purpose explicitness, purpose specificity, intelligible consent, consent with clear and plain language, freely given consent, and informed consent. After manually annotating 407 cookie banners, they found that 89% of cookie banners violated the European law, with 61% of those banners stating vague purposes and 30% using positive framing. This finding revealed a majority of cookie banners texts violate the GDPR's requirements of freely given and informed consent.

The language used in cookie banners is often formulated in a way that can confuse and impact users' privacy decisions, steering them to accept consent to tracking. Regulators, policymakers and scholars (Article 29 Working Party, 2018; CNIL, 2022; de l'Informatique et des Libertés, 2019; European Data Protection Board, 2022; Gray et al., 2018; European Data Protection Board, 2020), confirm that certain textual strategies such as the use of motivational language and humor (European Data Protection Board, 2022; Frobrukerrådet, 2018), shame (Mathur et al., 2019b), guilt (Brignull, 2010), blame (de l'Informatique et des Libertés, 2019), fear (Bongard-Blanchy et al., 2021) or uncertainty (European Data Protection Board, 2020) influence users' online decisions. Such textual expressions can violate the legal requirements for consent. Consent, if not obtained in compliance with the GDPR, provides invalid grounds for data processing, rendering the processing activity illegal (Article 6(1)(a), (EU, 2018))).

Chapter 3

Natural Language Processing

3.1 Natural Language Processing Architectures

Natural language processing (NLP) is a subfield within computer science, linguistics and artificial intelligence that is concerned with analysing and representing naturally occurring texts for the purpose of achieving human-like language processing. Natural language tasks can be split roughly into three categories: generation, classification and retrieval. We will focus on text classification in this project. Text classification is the process of categorising text into groups. Examples of text classification tasks include: sentiment analysis, topic detection, spam detection and language detection. Until very recently, general NLP tasks such as text classification were often managed with architectures based on word embeddings, machine learning, convolutional neural networks and recurrent neural networks.

3.1.1 Basics of NLP

Before describing the state-of-the-art NLP architectures, we will first go over some commonly used NLP methods and concepts.

Tokenisation Tokenisation is the process of segmenting text into 'tokens'. A piece of text can be tokenised by segmenting it into words and sentences, and sometimes the punctuation is also removed during this process.

Normalisation: stemming and lemmatisation Normalisation is the process of removing inflections from words. In order to do this, stemming and lemmatisation are used. Stemming is the process of removing the ends of words and includes the

removal of derivational affixes. Lemmatisation is the process of returning a lemma of a word: the base (or dictionary) form of a word.

NLP processes textual data with models, but the text needs to be converted into vectors before the textual data can be given as input to an algorithm. This process is called feature extraction. Commonly used techniques for feature extraction are Bag Of Words (BOW), N-grams, parts-of-speech (POS) tags and word embeddings.

Bag of Words Bag Of Words is a commonly used model that creates an occurrence matrix for all words in a piece of text. It represents a document vector by its word frequencies, disregarding word order and grammar. These word occurrences can then be used as features for training a classifier. The Bag Of Words model is mostly used in document classification methods, such as text classification (Soumya George & Joseph, 2014; VM, Kumar R, et al., 2019; Yan et al., 2020). Although the Bag Of Words method is considered a solid method to represent textual data, it has one major drawback. Since the number of features in the vectors increases significantly as the number of documents increases to account for all word occurrences, the dimension of the vectors can become tremendously large and sparse.

N-grams N-grams are continuous sequences of N words or tokens in a document. A 1-gram (unigram) is just one word or token, whereas a 2-gram (bigram) are two consecutive words or tokens, and so on. In the sentence "I am at home", there are four unigrams (I,am, at, home), three bigrams (I am, am at, at home), two trigrams (I am at, am at home) and one 4-gram (I am at home). N-grams and the probabilities of the occurrences of specific words in specific sequences can be used to predict language. N-grams are used in tasks like spelling correction, spam detection and text summarising (Aiyar & Shetty, 2018; Ashour et al., 2018; Chua & Asur, 2013; Ganesan et al., 2012; He et al., 2022).

POS-tags Parts of speech (POS) tags are a popular NLP method where labels are assigned to each word in a text that indicates the parts of speech. Commonly, this label also includes other grammatical aspects such as tense, number and case. The general structure of lexical terms within a phrase or text is described by POS-tags, so we can use them to infer semantics. Other applications of POS-tagging include named entity recognition, co-reference resolution and speech recognition (Aguilar et al., 2019; Cristea et al., 2002; Sun et al., 2021; Zuo et al., 2019).

Word Embeddings Word embeddings are learned representations of words in a numerical way using vectors. Each word is represented as a vector within a predefined

vector space. The distributed representation is based on the usage of the word, which allows words to have similar meanings with similar representations. Word embeddings can thus be trained and used to find relations and similarities between words. The benefits of word embeddings is that they can be derived from large unannotated corpora that are readily available and thus do not need extensive manual annotation. Two well known word embedding approaches are Word2Vec (Church, 2017) and GloVe (Pennington et al., 2014). Word embedding vectors can be used in natural language tasks such as sentiment analysis (Liu, 2017), document clustering (Mohammed et al., 2020) and speech tagging (Thavareesan & Mahesan, 2020). Drawbacks of this approach include the inability to handle unknown or out-of-vocabulary words and the inability to distinguish between different meanings of a word.

TF-IDF Term frequency-inverse document frequency, shortened to TD-IDF, is a statistical measure to quantify the importance of words or phrases to a document in a collection of documents. TD-IDF is commonly used in information retrieval and keyword extraction, and is also useful in machine learning algorithms for NLP. TF-IDF is calculated by multiplying the term frequency of a word in a document by its inverse document frequency in a collection of documents. The latter looks at how common the word is amongst the document set. The higher the TF-IDF score is, the more relevant the word is in a particular domain.

3.1.2 Machine Learning

Machine learning (ML) is a type of artificial intelligence that gets computers to act without being explicitly programmed. ML has been used in all types of applications and tasks, including NLP. Supervised machine learning infers a function from labelled training data (input-output pairs) and is then able to map this to new data. Popular machine learning algorithms include support vector machines (SVM) and Neural Networks (NN).

Support Vector Machines

A support vector machine can be used for regression and classification tasks. Although it is a simple machine learning algorithm, it is also one of the most robust prediction methods. In SVM data points are viewed as an n-dimensional vector (n being the number of features) in a space and the goal of

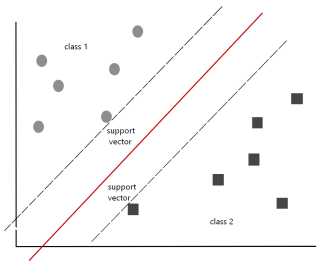


Figure 3.1: Support vector machine

the SVM is to find a (n-1)-dimensional hyperplane that separates the different classes of data points (see Figure 3.1). The hyperplane should have the maximum distance to the data points. The data points with the minimum distance to the hyperplane are known as support vectors and influence the position and orientation of the hyperplane. In their most simple form, SVM are used for binary classifications, dividing data into two distinct classes. If the output of the linear function is 1, it is identified as one class and if the output is -1, it is identified as the other class. SVM work relatively well when there is a clear space of separation between classes and when the number of dimensions is greater than the number of samples. The SVM are, however, less suitable for larger datasets or datasets that contain more noise. Although SVM have been quite successful in various text classification tasks (Chau & Chen, 2008; Sebastiani, 2002; Song et al., 2007; Yang et al., 2013), recently the attention in NLP has shifted towards state-of-the-art models like BERT (see Section 3.1.4). Research by Clavié and Alphonsus (2021) has shown that SVM classifiers perform surprisingly well on legal text classification and that there is a relatively small improvement between BERT and SVM within the legal domain.

3.1.3 Neural networks

Neural networks are very useful for various natural language tasks. The main motivation to use neural networks in NLP is to come up with more precise techniques than using word frequencies. Neural networks are computational nonlinear models, inspired by the neural structure of the brain. These neural networks consist of three interconnected layers: input layer, hidden layer(s) and output layer. Each node in the network has a weight and threshold or activation function. If the output from a node is above the specified threshold, the node is activated and the data is sent to the next layer of the network. Neural networks rely on training, which is the process of optimising the weights, to minimise the prediction error and improve the accuracy over time. A simple neural network is shown in Figure 3.2.

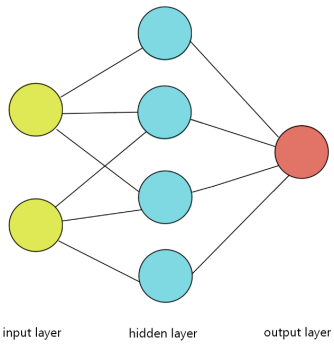


Figure 3.2: Simple neural network

CNN

Convolutional Neural Networks are a variant of neural network that contain one or more convolutional layers. These layers apply a convolution operation on the input and pass the data to the fully connected layer at the end of the network. The convolution operation is the process to detect the most important features from the input data. Although convolutional neural networks have primarily been applied in tasks related to computer vision, such as object detection and image classification, they also have been used for natural language processing tasks, such as sentiment analysis (Alharbi & de Doncker, 2019; Kalchbrenner et al., 2014; Ouyang et al., 2015). The constraints for CNNs include only accepting input and producing output in the form of a fixed-sized vector.

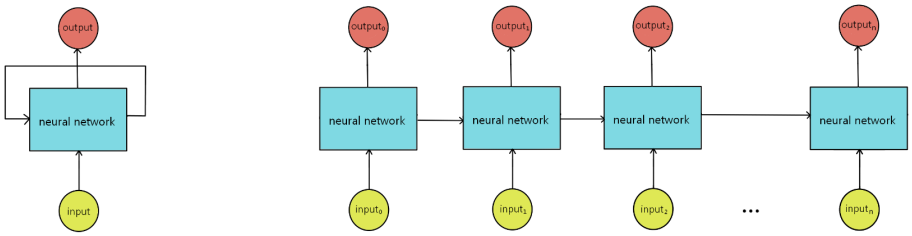


Figure 3.3: Left: recurrent neural network, right: recurrent neural network 'unfolding' in time

RNN

Unlike CNNs, recurrent neural networks allow for operation over sequences of vectors in the input and output. Moreover, RNNs are capable of working with varying sentence lengths, which cannot be achieved with a traditional neural network, and provides the additional benefit of learning features from different text positions. RNNs are a type of neural network in which nodes are connected in a directed cycle: a feedback loop. As a result, the output does not only depend on the present input, but also on previous input (Fig. 3.3). RNNs have been very effective in natural language generation (Bowman et al., 2015; Graves, 2013), machine translation (Cho et al., 2014) and speech recognition (Graves, Mohamed, et al., 2013).

LSTM

The output of a RNN might be passed into another RNN, or any number of layers of RNNs, to get more levels of computation to solve or approximate increasingly

complicated tasks. The increase in number of layers of RNNs introduces the vanishing gradient problem: the gradients become too close to zero, making the network impossible to train. A small gradient means that the weights are not updated effectively during training which can lead to inaccuracy of the neural network. Long Short Term Memory (LSTM) networks can be used to solve the vanishing gradient problem. LSTM is a small neural network that has four layers. One is the recurring layer from the RNN and the other three are networks that function as gates: an input gate, an output gate and a forget gate. The input gate controls what new information is encoded at each time step. The output gate controls how much information is sent to the next layer. The forget gate controls what information will be forgotten. The presence of activations in the forget gate enables the network to decide that certain information should not be lost, allowing the network to control the gradients' values more effectively and updating the parameters of the model suitably. RNNs with LSTM have shown promising results on speech recognition (Graves, Jaitly, et al., 2013; Sak et al., 2014), word segmentation (Yao & Huang, 2016) and sentence embedding (Palangi et al., 2015).

Attention-based Transformers

RNN-based architectures achieved state-of-the-art results in the past, but they are limited by their sequential nature when handling long text. Currently, attention-based transformers are gradually becoming the state-of-the-art in NLP. Vaswani et al. (2017) started the rise of the transformer model. This model was inspired by encoder-decoder architectures. Encoder-decoder models are a widely used subclass of seq2seq models, which are a class of models that transform a sequence into another sequence. An encoder-decoder consists of an encoder, a decoder and a hidden vector. The encoder converts input into a hidden vector and the decoder converts this hidden vector into output.

Transformers use attention techniques to capture information about a word's relevant context, which is subsequently encoded in a rich vector that intelligently depicts the word. An attention mechanism determines at each step which parts of an input sequence are important. So for each input that an encoder gets, the attention mechanism considers numerous other outputs at the same time and decides which ones are as important by assigning different weights to those inputs. The encoded sentence and the assigned weights will then be sent into the decoder. When the input consists of a very long sentence, it is difficult to capture all information. Attention mechanisms try to overcome this problem by allowing the decoder to access all the hidden states instead of just a single vector.

Transformers only use the attention mechanism rather than having a RNN to

encode each position. It has multiple layers of self-attention, which is an attention mechanism where the representation of a sequence is computed by relating different words in the same sequence. Transformers are trained self-supervised, which reduces the dependency on labelled input and permits the use of a larger pool of text, and they are very effective in transfer learning. The latter allows for pre-training them with large amounts of general-purpose texts and then to fine-tune them for their specialised tasks with good results, less effort, and less labelled data.

3.1.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a well-known example of a transformer model (Devlin et al., 2019). As opposed to directional models who read input sequentially, BERT is bidirectional, reading the entire sequence of words at once and learning the context of the word from all its surroundings (left and right of the word). The framework of BERT consists of two important steps: pre-training and fine-tuning.

Pre-training During pre-training, BERT is trained on unlabelled data in two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In the MLM task, 15% of input tokens are masked and then predicted. NSP is used in order to train a model that understands sentence relationships. The corpora used for pre-training are BooksCorpus and English Wikipedia.

Fine-tuning For fine-tuning, the model is first initialised with the pre-trained parameters, and all of the parameters are fine-tuned using labelled data from the downstream tasks. For each task, the task-specific inputs and outputs are put into BERT and the parameters are fine-tuned end-to-end.

3.1.5 Fine-tuned BERT

Pre-training language models is usually computationally expensive. This step can be skipped when using BERT, since the pre-trained models are publicly available. As BERT has been pre-trained on generic corpora, the models have under-performed in specialised domains. Chalkidis et al. (2020) explores the strategies to overcome this limitation in the legal domain by developing LEGAL-BERT. LEGAL-BERT are a family of BERT models that have been pre-trained on diverse English legal texts

from several fields, including European legislation (EURLEX[1]), UK legislation [2], and various courts proceedings from Europe and the US[3]. Results showed that both pre-training BERT on domain specific corpora and pre-training BERT from scratch on domain specific corpora had a better performance than using BERT out of the box, and were comparable in three legal datasets (Chalkidis et al., 2020).

3.1.6 Conclusion

As described, BERT is widely-used Transformer-based model, which serves as the basis for a variety of text classification tasks. The major advantage of BERT is that it was pre-trained on a large corpus, allowing it to be fine-tuned on a downstream task with a relatively small dataset. It thus seems as a suitable model to use for our task classification task.
While cookie banners are not themselves legal texts, they do explain legally relevant provisions; hence, we include the LEGAL-BERT model to address the utility of a domain-specific BERT model in the general legal domain.

3.2 Zero shot learning

The term "zero-shot learning" (ZSL) most frequently refers to a specific kind of task: training a classifier on one set of labels and evaluate it on a another set of labels that it has never seen before. Recently, getting a model to perform something that it wasn't specifically trained to do has become the more broad meaning of ZSL, notably in NLP (Sarkar et al., 2021; Ye et al., 2020; Yin et al., 2019b). This is also illustrated in Radford et al. (2019), where authors assess a language model on downstream tasks without directly fine-tuning on these tasks. The advantage of this machine learning technique is that it uses very few or even no labelled examples, which is very useful when there is only a small amount of data available for training. On the other hand, zero-shot requires descriptive and meaningful labels. Yin et al. (2019a) proposes using a pre-trained MNLI sequence-pair classifier as an out-of-the-box zero-shot text classifier. Natural language inference (NLI) considers a "premise" and a "hypothesis"

[1]Publicly available from http://eur-lex.europa.eu/

[2]Publicly available from http://www.legislation.gov.uk

[3]Cases from the European Court of Justice (ECJ), also available from EURLEX, cases from HUDOC, the repository of the European Court of Human Rights (ECHR) (http://hudoc.echr.coe.int/eng), cases from various courts across the USA, see https://case.law and US contracts from EDGAR, the database of US Securities and Exchange Commission (SECOM) (https://www.sec.gov/edgar.shtml).

to determine whether the hypothesis is true or false, given the premise. This can be adapted to zero-shot by making a sequence be the premise and turning a candidate label into the hypothesis. If the model predicts the hypothesis to be true, this can be taken as the prediction that the label applies to the sequence.

3.2.1 Conclusion

As mentioned, an advantage of zero-shot learning is that it requires few or no labelled examples. Since our dataset is relatively small, this approach can very useful. The out-of-the-box zero-shot classifier proposed by Yin et al. (2019a) seems to be an appropriate approach, since this is a simple but accurate method.

3.3 Stylistic classification

Style classification is a sub-field of text categorisation which is concerned with the aspects of linguistic expression rather than the text's content. Since the language used in cookie banners has particular linguistic styles and expressions, stylistic classification as an approach could be used with one of the architectures described in the previous subsections (Section 3.1). Classification of stylistic aspects occurs for example in authorship detection, deception detection or sentiment analysis. Hence, in this subsection we will describe stylistic classification in these task and go over different features used in stylistic classification.

3.3.1 Authorship detection

The task of verifying authorship of a text has been around for a long time, but had been gaining attention due to new application in forensic analysis and development of computational techniques (Koppel et al., 2009). Authorship classification commonly used multiple independent features, such as frequency of function words, N-grams, and word and sentence length statistics (Gamon, 2004). With the rise of machine learning techniques, text categorisation, and thus authorship detection, gained a new method to classify text, such as neural networks, k-nearest neighbours and support vectors machines (Koppel et al., 2009). The increase in complexity of techniques also leads to a decrease in the required text length to reach a good classification accuracy. Whereas previous work focused on longer documents, more recent work looks at shorter online content such as blog posts (Mohtasseb, Ahmed, et al., 2009) and Tweets (Layton et al., 2010).

3.3.2 Sentiment analysis

Sentiment analysis (SA) is concerned with extracting or classifying sentiment from reviews using various NLP and text analysis techniques. Analysing sentiment can be done on a document level, sentence level, word/term level or aspect level (Hussein, 2018). Important representations for sentiment analysis are bag-of-words, N-grams, TF-IDF, parts of speech tags and sentiment lexicons (Feldman, 2013; Hussein, 2018). Sentiment lexicons are a collection of words and their associated sentiment polarity.

3.3.3 Deception detection

Digital disinformation is reported to be a major risk of modern society (Howell et al., 2013). The work on automatic deception detection focuses on small manually build corpora to construct models to detect deceptive product reviews and news articles (Volkova & Jang, 2018). Research on deception detection mainly relies on language complexity, syntax, psycholinguistic signals, and biased language, in combination with machine learning models. These models include linguistic features such as parts of speech tags, n-grams and readability (Mihalcea & Strapparava, 2009; Pérez-Rosas & Mihalcea, 2015). Other features that are used for deception detection are Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) and connotations (Volkova & Jang, 2018), which provide insights about emotions or feelings that certain words invoke, among other things.

3.3.4 LIWC

So far, commonly used methods in style classification include parts of speech tags and various word frequency techniques as raw word count, N-grams or TF-IDF. Another interesting text analysis method is LIWC (Pennebaker et al., 2001). LIWC is a dictionary-based tool that calculates the percentage of words in a text that fall into 80 linguistic, psychological and topical categories. These categories range from standard linguistic dimensions (e.g. pronouns, prepositions, negations) and other grammar (e.g. common verbs. quantifiers, interrogatives) to psychological processes (e.g. social, biological, affective and cognitive processes) and personal concerns (e.g. work, leisure, religion, death) and it is also a possibility to add your own category. In research on misleading and deceptive language in political news, the most important LIWC features were punctuation and psychological features (Shrestha et al., 2020). Misleading news also has a higher frequency of psychological words such as personal concerns (death and religion-related) and social words (social, family-related words

and male and female related words). Studies on stylometric deception detection using LIWC usually have a classification accuracy of 70% (Tomas et al., 2022). Furthermore, vague words of the LIWC dictionary are, for example, *possible*, *some* and *perhaps* (cognitive processes: tentative category). Since the word list is composed from research on psychology, medicine and business and was originally designed to assess a individual's cognitive writing style. The list of vague terms of the LIWC might thus not perform well in identifying vagueness in cookie banner texts. In their research on vague decisions in constitutional court rulings, Sternberg (2018) expands the vague words from LIWC by using word embeddings to select new, legal domain-specific candidates for the dictionary.

3.3.5 Conclusion

Style classification is used in different tasks, such as authorship detection, deception detection or sentiment analysis, and with different methods. Commonly used methods for style classification include parts of speech tags and various word frequency techniques as raw word count, N-grams or TF-IDF. Another method mostly used in deception detection is LIWC, a dictionary-based text analysis tool. As mentioned before, LIWC might not perform well in identifying vagueness in cookie banner texts due to the nature of the language in cookie banners. However, it would still be interesting to use LIWC for the classes misleading language, framing and technical jargon. Moreover, we can manually add our own categories to LIWC for a more accurate classification.

Chapter 4

Data

In this section, we will describe the dataset we used. We will go into detail on the annotation classes, and their corresponding classification labels and legal violation. Lastly, we will also briefly explain what challenges are involved in this dataset of cookie banner texts.

4.1 Dataset

In Santos et al. (2021), cookie banner texts were manually annotated according to the GDPR legal requirements and their corresponding violations. The resulting dataset consists of 407 English cookie banner text segments. The full cookie banner texts have an average of 3.59 sentences and 49.77 words. The most common content words (i.e. 'cookies', 'website', 'policy', etc.) are very specific to the context of cookie banners.

4.2 Annotation classes and classification labels

The annotation classes and classification labels are based on the annotation guidelines used by the five experts for the study in Santos et al. (2021), where a given annotation *class* has one or more corresponding *labels*. The original dataset annotated texts segment-wise. In contrast, the goal of the present work was to label the cookie banner as a whole, to indicate whether it contains one or more instances of language that falls under any of these labels. The labels assigned to each cookie banner are thus determined by the presence of the labels in their text segments, in the original data. Thus, some segments might belong to more than one class and label.

Due to data sparseness, some classes in the original guidelines by Santos et al. (2021) were omitted, leaving five classes in total: *Consent options presence*, *Misleading language*, *Framing*, *Purpose* and *Technical jargon* (see Table 4.2).

As described by Santos et al. (2021), there are six legal requirements and their corresponding violations applicable to cookie banner texts (Table 4.1). There are thus multiple ways that cookie banner texts can violate legal requirements, such as using technical jargon or prolixity (violation of R3, R4.1), the absence of purpose or vague purposes (R1, R2), using positive or negative framing (R5), or using misleading vague language (R4). A more detailed explanation of possible legal violation in combination with the different classes is given in Section 2.

Legal requirement	**Violation**
R1 Purpose explicitness	
R1.1 Availability	Absence of purpose
R1.2 Unambiguity	Ambiguous intent
R1.3 Shared common understanding	Inconsistent purposes
R2 Purpose specificity	Vague or general purposes
R3 Intelligible consent	
R3.1 Non-technical terms	Presence of technical jargon
R3.2 Conciseness	Prolixity
R4 Consent with clear and plain language	
R4.1 Straightforward statements	Misleading expressions
R4.2 Concreteness	Indefinite qualifiers
R5 Freely given consent	Pressure to provide consent
R6 Informed consent	Absence of essential information about data processing

Table 4.1: Legal requirements and their corresponding violations

4.3 Legal violation and annotation

Here, we will brief describe which violation can occur in each class. In Table 4.2, each class is shown with their corresponding possible violation of legal requirements (based on Santos et al. (2021), see Table 4.1).

For consent options presence, not having an option to decline or reject to give consent to a website to use cookies violates the requirement to freely give consent

Class	Possible violation of legal requirement
Consent options presence	R5 Freely given consent
Framing	R5 Freely given consent
	R6 Informed consent
Misleading language	R1.2 Unambiguity
	R2 Purpose specificity
	R4 Consent with clear plain language
	R6 Informed consent
Purpose	R1.1 Availability
	R1.2 Unambiguity
	R2 Purpose specificity
Technical jargon	R3 Intelligible consent
	R4.1 Straightforward statements

Table 4.2: Annotation categories and the corresponding violations

(R5). Both negative and positive framing can highlight certain aspects of cookie use and purposes, which may nudge users towards giving their consent while not fully understanding what they consent to. This violates the requirements of freely given consent (R5) and informed consent (R6). Ambiguity and vagueness are misleading when users are uncertain about the intended meaning of the text, which can make the consent of the user uninformed. Misleading language can thus violate requirements of unambiguity (R1.2), specificity (R2), consent with clear plain language (R4) and informed consent (R6). The absence of purpose violates the requirement of purpose availability (R1.1), whereas a lot of different purposes (especially in a short part of the text, such as in one single sentence) can violate the requirement of purpose specificity (R2) and possibly unambiguity (R1.2). The presence of technical jargon breaches the requirement of intelligibility (R3) and straightforward statements (R4.1).

4.4 Challenges data

Regarding the aim of this project and the available data, there are several challenges. The cookie banner texts are most likely very similar to each other, which can make it more difficult to reach an accurate classification. Furthermore, cookie banner texts are relatively short and 407 annotated cookie banners is also a small dataset for NLP. The distribution of the data is not ideal, since some annotated classes have a lot more occurrences than other classes. Lastly, cookie banner text is different from 'regular' natural language, due to the text being formal, the text including many content words specific to the content of cookie banners, and the cookie banners possibly

containing dark patterns, which makes the text misleading or vague.
To overcome these challenges, we will compare the performance of different models and include some processing steps to extract certain features before loading the data into the models. Due a possibility that the formal text of cookie banners might be similar to legal text, we have added a model that performs well on data from a legal domain, to investigate if these models have a similar or higher accuracy than the models not fine-tuned on legal text.

Chapter 5

Method

In the section 3.1, we gave a detailed overview of the most commonly used basic and complex NLP concepts as well as models. In this section, we will describe in more detail which models we have decided to use for the purposes of our study. Before explaining our models, we first describe our date pre-processing step as well, since for some of our models this step is quite crucial. Lastly, we will also discuss the evaluation of the models.

5.1 Data preprocessing

In this subsection we will first explain the challenges of cookie banner classification that is due to the way we pre-process the data, namely instead of classifying segment-wise. we classify the whole cookie banner text. Second, we will delve into the specific pre-processing steps for the separate models.

As a results of presenting the data banner-wise instead of segment-wise, some segments might belong to more than one class and label. The main problem with this is that some classes have a very high label variation. For example, the class Consent options presence originally had seven labels: consent by continuing text, more info option, accept option, reject option, manage cookie or change settings options, close "x", manage via browser. Since cookie banners have often multiple of these consent options presence, having the data classified banner-wise resulted in more than 30 different combinations of the original seven consent options presence labels. Moreover, almost 20 combinations had only 1 or 2 occurrences. We thus had to choose the labels not only based on the legal violations (as described as Section 4.3), but also based on practicality. The resulting class labels can be seen in Table 5.2.

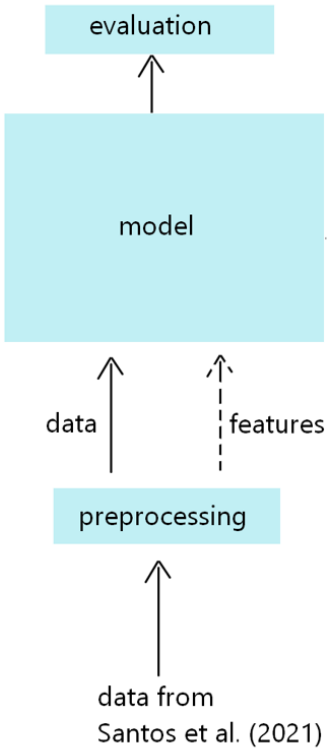


Figure 5.1: Visualization of the pipeline

To explain our choice in Table 5.2, here we briefly describe each main class, and the decisions made on how to combine various options into classifiable sub-classes.

Consent options presence. With the consent option presence class, we chose to make a distinction between whether there was a reject option in the cookie banner text or not.
Technical jargon. We kept the original annotation, which made the distinction on whether technical jargon was present or not.
Purpose. The original annotation of purpose consisted of eight labels: Essential functionalities, offering service, website/ux enhancement, profiling, advertisement, custom consent, analytics and social media features. Banner-wise data resulted in almost 40 different combinations of purpose labels. Due to this and an important legal violation being the absence of purpose, we chose the labels on whether a purpose was stated or not.
Framing. Framing labels originally included several types of framing: Best..., safety

Positive framing	**Negative framing**
Positive framing	Negative framing
Assumed happiness	Less functionalities
Safety or privacy arguments	Worse user experience
Compliance or authority argument	
Playful arguments	
Best...	

Table 5.1: Types of framing divided into negative and positive framing

or privacy arguments, less functionalities, assumed happiness, positive framing, compliance or authority arguments, negative framing, worse user experience, playful arguments. This resulted in more than 20 variations, with 14 combinations with 1 or 2 occurrences. Since these types of framing were originally divided into negative and positive framing (see Table 5.1), we decided to keep only this distinction, and added No framing to it. There were some instances of cookie banners that had both negative and positive framing. Since there we only a few, we manually looked at the text and labeled it based on which type of framing was more prominent.
Misleading language. Misleading language had four distinct labels: Misleading language, vagueness, deceptive language and prolixity. These labels combined into 10 different variations, with 3 combinations having only 1 or 2 occurrences. Due to only 3 instances of misleading language, we decided to remove these as a label. Furthermore, after initial testing of the labels showed a low accuracy, we decided to manually reduce the labels to the original main labels of vagueness, deceptive language and prolixity, similar to what we did with framing.

5.1.1 BERT with LIWC features

The complete LIWC uses 80 linguistic, topical and psychological categories. Before running BERT with LIWC, we had to decide which categories to use. Since the LIWC category 'tentative' from the cognitive processes category can be indicative for vague or deceptive language, we will use this category for the classification of misleading language.

We decided to manually add a new category based on the specific technical jargon found in our dataset, for classifying the class technical jargon. First, we manually looked at all the phrases annotated as technical jargon. Based on these phrases, we made a list of one-word terms that were frequently found in one or more technical jargon phrase. This was quite a task, since a lot of technical jargon phrases only occurred once and had little overlap in words or phrases. Moreover, most terms con-

Annotation class	Classification labels	Occurrences
Consent options presence	No reject option	344
	Reject option	63
Framing	No framing	239
	Positive framing	152
	Negative framing	16
Misleading language	No misleading language	267
	Vagueness	68
	Deceptive language	51
	Prolixity	21
Purpose	Purpose mentioned	328
	No purpose mentioned	79
Technical jargon	No technical jargon	331
	Technical jargon	76

Table 5.2: Annotation categories and classes

sisted of more than one word and often are in an expression where one word on itself cannot be deemed technical jargon, such as 'cookies and similar technologies', 'retract your compliance' or 'anonymised information will also be collected and processed'.

Some technical jargon words were excluded from the LIWC list, such as 'cookies', since these occur also frequently outside of technical jargon. The terms and some of their common technical jargon phrases are in Table 5.3. We made this list of terms into a new LIWC category. We also tested the full 80 LIWC categories for all classes.

Preliminary results showed that the BERT with the partial LIWC for misleading language and technical jargon had a decreased performance compared to BERT and BERT with full LIWC. We thus only used BERT in combination with the full 80 LIWC categories from here on.

5.1.2 BART in ZS-setting

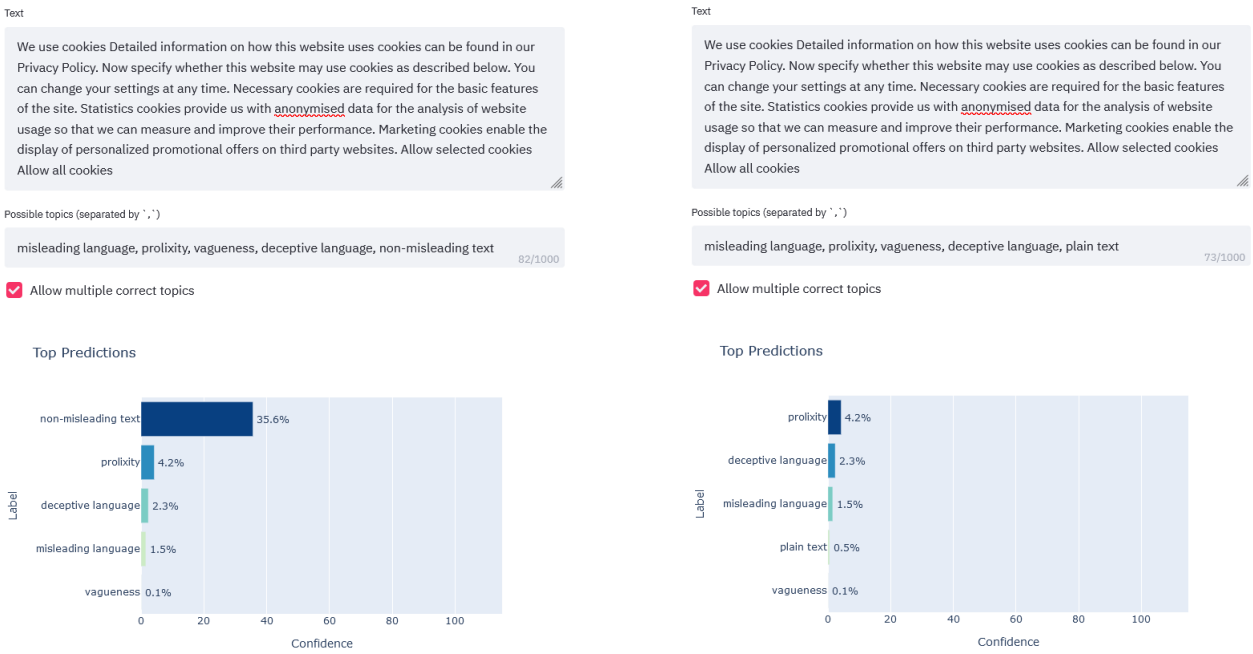
To use BART in ZS-setting required some specific pre-processing of the data as well. In this subsection, we will detail the steps taken towards that end.

For using BART in a zero-shot setting, it is essential to use descriptive and meaningful labels. Using the Hugging Face online zero-shot demo [1], we tested some cookie banner texts and possible labels to observe the predictions before running our own

[1]https://huggingface.co/zero-shot/

LIWC term	Technical jargon phrase
traffic	traffic
owner	domain owner
IP	IP adress, IP addresses
identifiers	cookie identifiers, identifiers assigned to the device
provider	service provider, third party provider
third-party	third-party cookies, third-party tools
party	third party, third party cookies
portal	web portal, this portal uses cookies
technical	technical cookies, technical and analytical cookies
aggregated	aggregated form

Table 5.3: Technical jargon LIWC terms and their common phrases



(a) Non-misleading text as label (b) Plain text as label

Figure 5.2: Zero-shot demo with a cookie banner example of misleading language

model on the complete data.

For the zero-shot labels, we use the original annotation labels when possible. The question is, however, what label to use for 'absence' of a particular annotation. For example, the original annotation of misleading language includes: misleading

language, deceptive language, prolixity and vagueness. The question for zero-shot classification of our data is how to annotate the absence of any type of, in this case, misleading language.

We use the zero-shot demo to how certain labels affect the classification confidence. As shown in Figure 5.2a and 5.2b, the wording of the exact labels does affect the classification. Using 'non-misleading text' as a possible topic leads to 35.6% confidence in this class, whereas this is only 0.5% when using 'plain text', and this also slightly affects the confidence in the other topics.

In general, terminology like 'non-misleading text' and 'no technical jargon' seem to prime the zero-shot model into over-classifying into this specific class, compared to labels like 'Neutral language'. So we mostly used terminology such as 'normal text/language' or 'other options' to achieve a more accurate classification. The final classification labels for BART in ZS-setting are shown in Table 5.4.

Annotation class	**BART-ZS labels**
Consent options presence	Reject option Other options
Framing	Negative framing Positive framing Neutral language
Misleading language	Deceptive language Misleading language Prolixity Vagueness Neutral language
Purpose	Purpose mentioned No purpose mentioned
Technical jargon	Technical jargon Neutral language

Table 5.4: Annotation categories and classes

5.2 Models

In this section we will describe the four models we have chosen to use for the classification of cookie banner texts.

5.2.1 BERT

Devlin et al. (2019) is a widely-used Transformer-based model, which serves as the basis for a variety of text classification tasks, including topic classification, and sentiment analysis. As described in Section 3.1.4, BERT is is used as the basis for a variety of text classification tasks. The major advantage of BERT is that it was pre-trained on a large corpus, allowing it to be fine-tuned on a downstream task with a relatively small dataset. It thus seems as a suitable model to use for our task classification task. A disadvantage is that BERT is not pre-trained specifically on cookie banner text, which is why we also included BERT models with extra features (Section 5.2.2) and LEGAL-BERT (Section 5.2.3).

5.2.2 BERT with LIWC features

Linguistic Inquiry and Word Count (LIWC) Pennebaker et al. (2001) is a dictionary-based text analysis tool with linguistic, psychological and topical categories. LIWC calculates the percentage of words from the cookie banner text that fall into each category and creates a vector of all these percentages. We concatenate BERT embeddings with a LIWC vector representing all 80 categories used by LIWC. The remaining architecture is the same with BERT. For classes like *framing*, *misleading language* and *technical jargon*, we expect that LIWC will increase the performance of the model, since these features reflect the more stylistic aspects of the text.

5.2.3 LEGAL-BERT

LEGAL-BERT are a family of BERT models that have been pre-trained on diverse English legal text from several fields. Since LEGAL-BERT model performs better than BERT on domain-specific tasks Chalkidis et al. (2020), we use the general LEGAL-BERT as a comparison for the BERT model. While cookie banners are not themselves legal texts, they do explain legally relevant provisions; hence, we include this model to address the utility of a domain-specific BERT model in the general legal domain.

5.2.4 BART in ZS-setting

Zero-shot (ZS) classification in NLP has been used to classify text on which a model is not specifically trained. As mentioned, an advantage of zero-shot learning

is that it requires few or no labelled examples. Since our dataset is relatively small, this approach can very useful. The out-of-the-box zero-shot classifier proposed by Yin et al. (2019a) seems to be an appropriate approach, since this is a simple but accurate method. So, we use the pre-trained BART-Large MNLI model Lewis et al. (2019) as an out-of-the-box zero-shot text classifier. Here, the cookie banner text is the premise and the corresponding labels are hypotheses. We use the model to estimate the probability of each label for every cookie banner text segment. The label with the highest probability is selected.

5.2.5 Training details and hyperparameters

For simplicity, a separate model was trained for each class.
For the fine-tuned models based on BERT and BERT-LEGAL, we use a classification layer of size 768, followed by a ReLU layer, to determine the most probable label for each class.
For BERT and BERT+LIWC features, we use BERT Base-cased. Since Base-Cased is not available for LEGAL-BERT, we use LEGAL-BERT Base-uncased. For the BERT-like models, the learning rate is set as 1e-6, the model is trained by using cross-entropy loss and the Adam optimizer.
The pre-trained models are loaded and fine-tuned on our embedded training data. The training was set for 12 epochs. For reporting our results, we used a 5-fold cross-validation setup. As our dataset is small and the class distributions are not balanced, we preferred a stratified split. Since BART is used in a zero shot-setting, cross-validation is not applicable for this model, and the results are reported accordingly.
All of the models were run on a laptop with AMD Ryzen 7 5700U processor (1.80 GHz) and 16 GB DDR 4 RAM.

5.3 Evaluation

All models will have an classification accuracy score, calculated as the average out of the stratified 5-fold cross-validation sets, as well as F1-scores for all class labels. Furthermore, we will use McNemar's test for model comparison.

Chapter 6

Results

In this section, we discuss the evaluation results of the models. First, we will review the accuracy and F1-scores per class. Thereafter, we compare the overall performance of the models and discuss the results of the McNemar's test. Lastly, we provide some classification examples and discuss the occurrence distribution.

Table 6.1 shows the performance of these models in terms of the average classification accuracy, computed as a proportion of correctly labelled instances per class. We provide F1-scores for all classes in Table 6.2. The accuracy scores and F1-scores per cross-validation set can be found in the Appendix, Section 9.2.

Accuracy performance differs for each class. Overall, we do not have a model that outperforms all the others for all classes, as the best accuracy performance for each class differs. However, LEGAL-BERT produces the best accuracy scores for four out of the five classes: Framing, Misleading language, Purpose and Technical jargon. BERT has the highest accuracy for the remaining class Consent options presence.

Baseline accuracy. In Table 6.1, we have added a majority baseline accuracy score for each class, based on the label that has the most occurrences per class. The baseline score is the accuracy score if all occurrences were classified as the majority label of the class. In Table 6.2, the labels and their occurrences are shown for all classes.

Consent options presence: The accuracy percentage is high for all models, but the highest score is from BERT with 92.9%, which is only small difference with the scores from BERT+LIWC and BART-ZS. All models have a higher accuracy than the baseline. The F1-scores are high for the majority label, and also quite high for the minority label with the exception for LEGAL-BERT.

Class	Baseline	BERT	BERT+LIWC	LEGAL-BERT	BART-ZS
Consent options presence	84.5	**92.9 (±1.77)**	92.1 (±1.65)	87.0 (±1.85)	91.65
Framing	58.7	71.0 (±2.15)	63.7 (±4.07)	**73.9 (±4.35)**	58.23
Misleading language	65.6	63.7 (±3.41)	60.2 (±5.25)	**66.1 (±0.51)**	54.30
Purpose	80.6	89.7 (±2.97)	90.9 (±2.28)	**92.9 (±2.49)**	76.90
Technical jargon	81.3	76.9 (±1.47)	75.2 (±2.38)	**80.3 (±2.13)**	78.87

Table 6.1: Comparison of cross-validation accuracy (mean and std) with best score per class in bold

Misleading language and *Framing*: These labels have the lowest accuracy percentages out of the five classes, with accuracy percentages dropping to 60% for some models. We also observe the lowest occurrences in these classes, with very low or null F1-scores. Given that these are the classes with more than two labels and rely on stylistic aspects of the text, these results are not surprising.
Misleading language: LEGAL-BERT has the best score with 66.1%, which is also the only score that is higher than the baseline. BART has the lowest score, even dropping below 55%. The Prolixity label has null F1-scores for all models except BERT+LIWC.
Framing: LEGAL-BERT produces the highest accuracy score for Framing with 73.9%. BART has the lowest accuracy, even dropping just below the baseline accuracy. The Negative Framing label has null F1-scores for all models except BERT+LIWC and BART-ZS.

Purpose: The highest accuracy comes from LEGAL-BERT with 92.9%, with BERT-LIWC still high with 90.9%. In general, this class suffers the least from the overfitting of the majority label, and has overall higher F1-scores for both labels. BART-ZS performs the worst with 76.9%, the only model below the baseline score, and has the lowest F1-scores.

Technical jargon: Interestingly, all models' scores are below the baseline score of 81.3%. LEGAL-BERT gives the best result with 80.3%, with all other models' score sitting in the 75-80% range. In general F1-scores are high for the majority labels and not the minority labels, but this is especially the case for LEGAL-BERT with only an F1-score of 0.04 for the minority label.

Model comparison. Overall, LEGAL-BERT and BERT perform well. LEGAL-BERT has the highest accuracy for four out of five classes. BERT, on the other hand, has the highest accuracy for one class and higher F1-scores for minority classes, compared to LEGAL-BERT. BERT+LIWC is similar to BERT: accuracy scores and F1-scores are similar, as well as no different proportion of errors between the two mod-

Class	Label	Occ. total	BERT	BERT+LIWC	LEGAL-BERT	BART-ZS
Consent options	Other	344	0.96 (±0.01)	0.95 (±0.01)	0.93 (±0.01)	0.95
presence	Reject option	63	0.73 (±0.06)	0.70 (±0.06)	0.38 (±0.22)	0.68
Framing	No framing	239	0.79 (±0.01)	0.71 (±0.04)	0.80 (±0.04)	0.73
	Positive	152	0.61 (±0.05)	0.57 (±0.05)	0.68 (±0.07)	0.17
	Negative	16	0.00 (±0.00)	0.04 (±0.09)	0.00 (±0.00)	0.13
Misleading	None	267	0.79 (±0.03)	0.78 (±0.04)	0.82 (±0.01)	0.71
language	Vagueness	68	0.19 (±0.08)	0.21 (±0.13)	0.13 (±0.12)	0.16
	Deceptive lang.	51	0.23 (±0.19)	0.23 (±0.19)	0.11 (±0.13)	0.04
	Prolixity	21	0.00 (±0.00)	0.04 (±0.09)	0.00 (±0.00)	0.00
Purpose	Yes	328	0.94 (±0.02)	0.94 (±0.01)	0.96 (±0.01)	0.87
	None	79	0.71 (±0.09)	0.75 (±0.07)	0.79 (±0.08)	0.00
Technical jargon	None	331	0.87 (±0.01)	0.85 (±0.02)	0.89 (±0.01)	0.88
	Yes	76	0.13 (±0.12)	0.16 (±0.10)	0.04 (±0.09)	0.09

Table 6.2: F1-results (mean and std) per classification label for all models

els. BART-ZS performs well for Consent options presence, but on all other classes the model's accuracy scores are below the majority baseline. Although BERT+LIWC outperforms BART-ZS on accuracy, the models do not have a different proportion of errors, except for the Purpose class.

To compare the classification results of the models, we used pairwise McNemar's tests, see Table 6.3. Overall, LEGAL-BERT and BERT achieved the highest scores. However, LEGAL-BERT's F1-scores are lower than BERT for minority classes. Comparing the two models with a McNemar's test we observe that they perform significantly differently for all classes, meaning that the models have a different proportion of errors. Looking at the result from McNemar's test for the other models, we see that BERT+LIWC/LEGAL-BERT and LEGAL-BERT/BART-ZS also have different proportions of errors. For BERT+LIWC/BERT and BERT+LIWC/BART-ZS, most classes are not significantly different. BERT/BART-ZS only have different proportions of errors for Framing and Purpose.

Class	BERT / BERT+LIWC	BERT / LEGAL-BERT	BERT BART-ZS	BERT+LIWC / LEGAL-BERT	LEGAL-BERT / BART-ZS	BERT+LIWC / BART-ZS
Consent opt. presence	.629	.000**	.583	.000**	.000**	.896
Framing	.002*	.000**	.000**	.000**	.000**	.129
Misleading language	.125	.000**	.011*	.000**	.000**	.126
Purpose	.56	.000**	.000**	.000**	.000**	.000**
Technical jargon	.371	.000**	.551	.000**	.000**	.248

Table 6.3: P-values of McNemar's test on all model combinations. $^{*}p < .05$, $^{**}p < .001$

Occurrence distribution: Studying the classes, and their corresponding misclassifications and the F1-scores, we observe that the data distribution affects the accuracy. Classification labels that have a low amount of occurrences in the data are almost

always incorrectly classified, even after the application of a stratified split for training and validation (see Table 6.2).

Observations. We provide some examples of (in)correct classifications of certain classes for all models, see Table 6.4. The corresponding cookie banner text segments are as follows:

1. In order to give you a better service our website uses cookies. By continuing to browse the site you are agreeing to our use of cookies. Further information. Yes, I agree.

2. This website or its third-party tools use cookies, which are necessary to its functioning and required to achieve the purposes illustrated in the cookie policy. If you want to know more or withdraw your consent to all or some of the cookies, please refer to the cookie policy. By closing this banner, scrolling this page, clicking a link or continuing to browse otherwise, you agree to the use of cookies.

3. We use cookies on this site to enhance your user experience Please read our Cookie policy for more info about our use of cookies and how you can disable them. By clicking the "I accept" button, you consent to the use of these cookies. More info I accept I do not accept.

4. This website uses cookies to enable you to place orders and to give you the best browsing experience possible. By continuing to browse you are agreeing to our use of cookies. Full details can be found here.

5. By using this site you agree to store cookies for the best site experience. More info Sure!

Banner text	Ground truth	BERT	BERT+LIWC	LEGAL-BERT	BART-ZS
1	**No framing**	No framing	No framing	Pos. framing	No framing
2	**Negative framing**	No framing	No framing	No framing	Positive framing
3	**Positive framing**	No framing	Pos. framing	No framing	Positive framing
1	**Vagueness**	Vagueness	No mislead. lang.	No mislead. lang.	Vagueness
3	**No mislead. lang.**	No mislead. lang.	Vagueness	No mislead. lang.	Vagueness
4	**Deceptive lang.**	No mislead. lang.	No mislead. lang.	No mislead. lang.	Deceptive. lang
2	**Techn. jargon**	No techn. jargon	No techn. jargon	No techn. jargon	No techn. jargon
3	**No techn. jargon**	No techn. jargon	No techn. jargon	No techn. jargon	No techn. jargon
3	**Purpose ment.**	Purpose ment.	Purpose ment.	Purpose ment.	Purpose ment.
5	**No purpose ment.**	Purpose ment.	No purpose ment.	Purpose ment.	Purpose ment.
2	**No reject opt.**	No reject opt.	No reject opt.	No reject opt.	No reject opt.
3	**Reject opt.**	Reject opt.	Reject opt.	No reject opt.	No reject opt.

Table 6.4: Example cookie banner text segments and their corresponding classification for each model

Chapter 7

Discussion

In this section, we discuss the interpretations and implications of the results from the previous section. We also describe surprising results and how these can be explained. Moreover, we will also briefly describe what can be improved upon and what would be useful to be included in future research.

As mentioned in Section 5.2.3, we included LEGAL-BERT to address the utility of a domain-specific BERT model in the general legal domain. LEGAL-BERT producing the highest score for most classes is interesting, since cookie banners themselves are not legal texts. The fact that they do explain legally relevant provisions might be the reason that this model has such a high classification accuracy. Cookie banner text classification is challenging, since the texts are short and the most common content words are very specific to the context of cookie banners. Further research on the language used in cookie banners, including whether legal text is similar to cookie banner text, could yield results that give more insight into which models are suitable for automatic classification of cookie banner texts.

One of advantages of using the BART-ZS model is that it requires few or no labelled examples, which is especially interesting for our small and unbalanced dataset. Our results from BART-ZS, however, were quite disappointing. The accuracy scores for all but one class were below the majority baseline, and the F1-scores for minority classes were low. These results are most likely due to a main practical constraint of the model, namely that it requires descriptive and meaningful labels. In Section 5.1.2, we explained our process of selecting labels to use for BART-ZS. Although we tried to make these labels as descriptive and meaningful as possible, most labels consisted of an original annotation label, whether something is present (e.g. 'technical jargon', 'deceptive language'), and an 'opposite label', whether something is not present (e.g. 'neutral language', 'other options'). These opposite labels are not that descriptive

and these kind of labels might not be entirely suitable for this model. An improvement could therefore be to come up with more meaningful and descriptive labels for these labels.

Another interesting result was the accuracy for the class Technical jargon. The majority baseline accuracy score was higher than the accuracy scores of all models. This might be due to the variation in technical jargon annotation: most text segments that were annotated as technical jargon by Santos et al. (2021) occurred only once. In addition, the most common technical jargon annotation was 'cookies' with 27 occurrences, which occurred in the majority—if not in all—cookie banners. This means that it occurred a lot more in the dataset without it being annotated as technical jargon, compared to the 27 times it was annotated as technical jargon. The inconsistency of the technical jargon annotation could be improved by having clear descriptions of when text is considered 'technical jargon' vs when it is not, as in the case of 'cookies'. This can give more insight to improve both the annotation and classification of this class.

Although LEGAL-BERT and BERT perform well, the results show that the data distribution affects the accuracy: minority labels are almost always incorrectly classified. In addition to the application of a stratified split for training and validation that we did, future research on small datasets should include more methods to work with such an imbalanced data distribution. For instance, weights could be added to minority classes to achieve a more balanced accuracy, or the unweighted average recall can be used as a better metric to optimise when the sample class ratio is skewed. Furthermore, to overcome the imbalanced data distribution as a whole, more data should be collected and annotated to achieve balanced data for all classes. Since manual annotation is very time-consuming and requires extensive expert knowledge, research should also include methods such as data augmentation to speed up data annotation.

In this thesis, we further add to the limited amount of studies on automatic detection of textual legal violations of cookie banners and lay a foundation for further research on this topic. Since the language and style of the cookie banners change rapidly, we need robust algorithms that can adapt to changes both in the legal domain and in the manner of adoption of new regulations by website operators. Hence, it is crucial to develop an efficient annotation pipeline to speed up human-in-the-loop annotation and automatic classification.

Chapter 8

Conclusion

In this thesis, we used a cookie banner dataset previously annotated by five experts who detected legal violations. First, we looked at which state-of-the-art deep learning models suitable to use for classification of legal violations, and selected three models: BERT, LEGAL-BERT and BART in a zero-shot setting. We also combined a dictionary based approach, i.e. LIWC embeddings with BERT, and checked whether this would improve performance.

Our approach aimed to give more insight into automatic detection of legal violations in cookie banner texts by comparing the performance of these four models. Our results suggest that there is not one model that outperforms all the others for all classes that needs to be detected. LEGAL-BERT works well in general for four out of five classes, but has lower F1-scores for minority classes. BERT also performs well, with the highest score for the remaining class and overall high scores on other class. BERT also has higher F1-scores for minority classes, compared to LEGAL-BERT. However, a close look reveals that the model is affected by the skewed data distribution for certain classes. In contrast, BART-ZS performs the worst for most of the classes, but it is not affected by the small size of the dataset, and the unbalanced distribution of the classes.

The results from this research show that using a state-of-the-art classification model off the shelf or with minimal fine-tuning will not yield reliable results for auditing or helping policymakers, since even the best performing models are affected by skewed data. Our initial tests give insight into which model performs well for which challenges, and can be used to build an efficient automatic classification pipeline of cookie banner texts in the future.

Chapter 9

Appendix

9.1 Ethical implications and limitations

In this paper, we rely on large, pre-trained language models for classification, fine-tuning them on a small, manually labelled dataset.

One limitation of this approach is the limited size of the manually labelled data. While accuracy and F1 figures may suggest reasonable performance on certain classes, we cannot consider such results as final, or as indicating that the models we use are sufficiently robust to be deployed in real-world settings. Rather, the results provide a picture of what current language models can achieve in a relatively under-explored domain, and provide directions for future work. As noted in the concluding section, one important direction is to curate larger and more diverse training data for the task of cookie banner classification.

9.2 Full results

Table 9.1, 9.2 and 9.3 show the accuracy scores for all cross validation sets for BERT, BERT with LIWC and LEGAL-BERT, respectively.

Class	**Set 1**	**Set 2**	**Set 5**	**Set 4**	**Set 5**	**Average**
Consent options presence	91.5	90.2	95.1	93.8	93.8	92.88 (±1.77)
Framing	68.3	74.4	69.1	71.6	71.6	71.00 (±2.15)
Misleading language	58.8	64.6	69.1	64.2	61.7	63.68 (±3.41)
Purpose	91.5	87.8	85.2	93.8	90.1	89.68 (±2.97)
Technical jargon	76.8	78.0	79.0	75.3	75.3	76.88 (±1.47)

Table 9.1: BERT

Class	**Set 1**	**Set 2**	**Set 5**	**Set 4**	**Set 5**	**Average**
Consent options presence	89.0	93.9	92.6	92.6	92.6	92.14 (±1.65)
Framing	65.9	57.3	67.9	66.7	60.5	63.66 (±4.07)
Misleading language	62.2	61.0	67.9	58.0	51.9	60.20 (±5.25)
Purpose	90.2	89.0	91.4	88.9	95.1	90.92 (±2.28)
Technical jargon	75.6	74.4	79.0	75.3	71.6	75.18 (±2.38)

Table 9.2: LIWCBERT

Class	**Set 1**	**Set 2**	**Set 5**	**Set 4**	**Set 5**	**Average**
Consent options presence	89.0	86.6	88.9	86.4	84.0	86.98 (±1.85)
Framing	76.8	80.5	71.6	72.8	67.9	73.92 (±4.35)
Misleading language	65.9	65.9	66.7	66.7	65.4	66.12 (±0.51)
Purpose	91.5	92.7	92.6	97.5	90.1	92.88 (±2.49)
Technical jargon	82.9	80.5	76.5	80.2	81.5	80.32 (±2.13)

Table 9.3: cv LEGALBERT

Table 9.4, 9.5 and 9.6 show the F1 scores per class label for all cross validation sets for BERT, BERT with LIWC and LEGAL-BERT, respectively.

Class	**Label**	**Set 1**	**Set 2**	**Set 5**	**Set 4**	**Set 5**	**Average**
Consent options presence	Other	0.95	0.94	0.97	0.96	0.96	0.96 (±0.01)
	Reject option	0.70	0.64	0.80	0.76	0.76	0.73 (±0.06)
Framing	No framing	0.76	0.80	0.78	0.80	0.79	0.79 (±0.01)
	Positive	0.58	0.70	0.57	0.57	0.63	0.61 (±0.05)
	Negative	0.00	0.00	0.00	0.00	0.00	0.00 (±0.00)
Misleading language	None	0.73	0.80	0.82	0.81	0.77	0.79 (±0.03)
	Vagueness	0.28	0.09	0.24	0.25	0.10	0.19 (±0.08)
	Deceptive lang.	0.33	0.43	0.40	0.00	0.00	0.23 (±0.19)
	Prolixity	0.00	0.00	0.00	0.00	0.00	0.00 (±0.00)
Purpose	Yes	0.95	0.93	0.90	0.96	0.94	0.94 (±0.02)
	None	0.74	0.58	0.67	0.85	0.71	0.71 (±0.09)
Technical jargon	Yes	0.00	0.31	0.00	0.17	0.17	0.13 (±0.12)
	None	0.87	0.87	0.88	0.86	0.86	0.87 (±0.01)

Table 9.4: F1 BERT

Class	Label	Set 1	Set 2	Set 5	Set 4	Set 5	Average
Consent options presence	Other	0.94	0.96	0.96	0.96	0.96	0.95 (±0.01)
	Reject option	0.61	0.80	0.67	0.73	0.70	0.70 (±0.06)
Framing	No framing	0.74	0.65	0.77	0.72	0.69	0.71 (±0.04)
	Positive	0.58	0.52	0.58	0.65	0.50	0.57 (±0.05)
	Negative	0.22	0.00	0.00	0.00	0.00	0.04 (±0.09)
Misleading language	None	0.79	0.78	0.84	0.76	0.71	0.78 (±0.04)
	Vagueness	0.18	0.21	0.30	0.38	0.00	0.21 (±0.13)
	Deceptive lang.	0.13	0.40	0.00	0.15	0.19	0.17 (±0.13)
	Prolixity	0.00	0.00	0.22	0.00	0.00	0.04 (±0.09)
Purpose	Yes	0.94	0.93	0.95	0.93	0.97	0.94 (±0.01)
	None	0.71	0.77	0.72	0.67	0.87	0.75 (±0.07)
Technical jargon	Yes	0.29	0.09	0.19	0.00	0.21	0.16 (±0.10)
	None	0.85	0.85	0.88	0.86	0.83	0.85 (±0.02)

Table 9.5: F1 LIWCBERT

Class	Label	Set 1	Set 2	Set 5	Set 4	Set 5	Average
Consent options presence	Other	0.94	0.92	0.94	0.93	0.91	0.93 (±0.01)
	Reject option	0.57	0.59	0.47	0.27	0.00	0.38 (±0.22)
Framing	No framing	0.82	0.85	0.79	0.79	0.74	0.80 (±0.04)
	Positive	0.72	0.79	0.60	0.68	0.61	0.68 (±0.07)
	Negative	0.00	0.00	0.00	0.00	0.00	0.00 (±0.00)
Misleading language	None	0.83	0.82	0.82	0.81	0.80	0.82 (±0.01)
	Vagueness	0.31	0.00	0.00	0.12	0.20	0.13 (±0.12)
	Deceptive lang.	0.00	0.25	0.29	0.00	0.00	0.11 (±0.13)
	Prolixity	0.00	0.00	0.00	0.00	0.00	0.00 (±0.00)
Purpose	Yes	0.95	0.95	0.96	0.98	0.94	0.96 (±0.01)
	None	0.74	0.81	0.75	0.94	0.73	0.79 (±0.08)
Technical jargon	Yes	0.22	0.00	0.00	0.00	0.00	0.04 (±0.09)
	None	0.90	0.89	0.87	0.89	0.90	0.89 (±0.01)

Table 9.6: F1 LEGAL-BERT

Bibliography

Aguilar, G., Maharjan, S., López-Monroy, A. P., & Solorio, T. (2019). A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135*.

Aiyar, S., & Shetty, N. P. (2018). N-gram assisted youtube spam comment detection. *Procedia computer science*, *132*, 174–182.

Alharbi, A. S. M., & de Doncker, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Cognitive Systems Research*, *54*, 50–61.

Article 29 Working Party. (2018). *Guidelines on transparency under regulation 2016/679, (wp260)* (tech. rep.).

Ashour, M., Salama, C., & El-Kharashi, M. W. (2018). Detecting spam tweets using character n-gram features. *2018 13th International conference on computer engineering and systems (ICCES)*, 190–195.

Bollinger, D., Kubicek, K., Cotrini, C., & Basin, D. (2022). Automating cookie consent and gdpr violation detection. *31st USENIX Security Symposium (USENIX Security 22)*.

Bongard-Blanchy, K., Rossi, A., Rivas, S., Doublet, S., Koenig, V., & Lenzini, G. (2021). "i am definitely manipulated, even when i am aware of it. it's ridiculous!" - dark patterns from the end-user perspective. *Proceedings of ACM DIS Conference on Designing Interactive Systems*. https://doi.org/10.1145/3461778.3462086

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proc. Priv. Enhancing Technol.*, *2016*(4), 237–254.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Brignull, H. (2010). Dark patterns [https://www.darkpatterns.org].

Brignull, H., Miquel, M., Rosenberg, J., & Offer, J. (2015). *Dark patterns - user interfaces designed to trick people*. http://darkpatterns.org/

Brownlee, J. (2016). *Why dark patterns won't go away*. https://www.fastcompany.com/3060553/why-dark-patterns-wont-go-away

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, *44*(2), 482–494.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chua, F., & Asur, S. (2013). Automatic summarization of events from social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 81–90.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Clavié, B., & Alphonsus, M. (2021). The unreasonable effectiveness of the baseline: Discussing svms in legal text classification. *arXiv preprint arXiv:2109.07234*.

CNIL. (2022). Deliberation of the restricted committee No. SAN-2021-024 of 31 December 2021 concerning FACEBOOK IRELAND LIMITED [https://www.cnil.fr/sites/default/files/atoms/files/deliberation_of_the_restricted_committee_no._san-2021-024_of_31_december_2021_concerning_facebook_ireland_limited.pdf].

Cristea, D., Postolache, O.-D., Dima, G.-E., & Barbu, C. (2002). Ar-engine-a framework for unrestricted co-reference resolution. *LREC*.

Curley, A., O'Sullivan, D., Gordon, D., Tierney, B., & Stavrakakis, I. (2021). The design of a framework for the detection of web-based dark patterns.

de l'Informatique et des Libertés, C. N. (2019). Shaping choices in the digital world [https://linc.cnil.fr/sites/default/files/atoms/files/cnil_ip_report_06_shaping_choices_in_the_digital_world.pdf].

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2018). We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020). Ui dark patterns and where to find them: A study on mobile applications and user

perception. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Article 29 Working Party. (2012). *Opinion 04/2012 on cookie consent exemption (WP 194)* (tech. rep.) [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp194_en.pdf].

EU, E. U. (2009). Directive 2009/136/ec of the european parliament and of the council of 25 november 2009 amending directive 2002/22/ec.

General Data Protection Regulation (2018). Retrieved January 31, 2022, from https://gdpr-info.eu/

European Data Protection Board. (2022). Guidelines 3/2022 on Dark patterns in social media platform interfaces: How to recognise and avoid them Version 1.0 Adopted on 14 March 2022 [https://edpb.europa.eu/system/files/2022-03/edpb_03-2022_guidelines_on_dark_patterns_in_social_media_platform_interfaces_en.pdf].

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89.

Frobrukerrådet. (2018). Deceived by design: How tech companies use dark patterns to discourage us from exercising our rights to privacy [https://www.forbrukerradet.no/undersokelse/no-undersokelsekategori/deceived-by-design].

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 611–617.

Ganesan, K., Zhai, C., & Viegas, E. (2012). Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. *Proceedings of the 21st international conference on World Wide Web*, 869–878.

Goodstein, S. A. (2021). When the cat's away: Techlash, loot boxes, and regulating" dark patterns" in the video game industry's monetization strategies. *U. Colo. L. Rev.*, *92*, 285.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. *2013 IEEE workshop on automatic speech recognition and understanding*, 273–278.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649.

Gray, C. M., Chivukula, S. S., & Lee, A. (2020). What kind of work do" asshole designers" create? describing properties of ethical concern on reddit. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 61–73.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of ux design. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Hausner, P., & Gertz, M. (2021). Dark patterns in the interaction with cookie banners. *arXiv preprint arXiv:2103.14956*.

He, J.-W., Jiang, W.-J., Chen, G.-B., Le, Y.-Q., & Ding, X.-F. (2022). Enhancing n-gram based metrics with semantics for better evaluation of abstractive text summarization. *Journal of Computer Science and Technology*, *37*(5), 1118–1133.

Howell, L., et al. (2013). Digital wildfires in a hyperconnected world. *WEF report*, *3*(2013), 15–94.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, *30*(4), 330–338.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kampanos, G., & Shahandashti, S. F. (2021). Accept all: The landscape of cookie banners in greece and the uk.

Khandelwal, R., Nayak, A., Harkous, H., & Fawaz, K. (2022). Cookieenforcer: Automated cookie notice analysis and enforcement. *arXiv preprint arXiv:2204.04221*.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, *60*(1), 9–26.

Layton, R., Watters, P., & Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. *2010 Second Cybercrime and Trustworthy Computing Workshop*, 1–8.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, H. (2017). Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*.

Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, *13*(1), 43–109.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019a). Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–32.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019b). Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–32.

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What makes a dark pattern... dark? design attributes, normative considerations, and measurement methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18.

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 309–312.

Mohammed, S. M., Jacksi, K., & Zeebaree, S. R. (2020). Glove word embedding and dbscan algorithms for semantic document clustering. *2020 International Conference on Advanced Science and Engineering (ICOASE)*, 1–6.

Mohtasseb, H., Ahmed, A., et al. (2009). Mining online diaries for blogger identification.

Narayanan, A., Mathur, A., Chetty, M., & Kshirsagar, M. (2020). Dark patterns: Past, present, and future: The evolution of tricky user interfaces. *Queue*, *18*(2), 67–92.

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020). Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015). Sentiment analysis using convolutional neural network. *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2359–2364. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349

Palangi, H., Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R. (2015). Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval. arxiv. org.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1120–1125.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

Santos, C., Rossi, A., Sanchez Chamorro, L., Bongard-Blanchy, K., & Abu-Salma, R. (2021). Cookie banners, what's the purpose? analyzing cookie banner text through a legal lens. *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, 187–194.

Sarkar, R., Ojha, A. K., Megaro, J., Mariano, J., Herard, V., & McCrae, J. P. (2021). Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. *Proceedings of the Natural Legal Language Processing Workshop 2021*, 102–106.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1–47.

Shrestha, A., Spezzano, F., & Gurunathan, I. (2020). Multi-modal analysis of misleading political news. *Multidisciplinary International Symposium on Disinformation in Open Online Media*, 261–276.

Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (2020). Circumvention by design-dark patterns in cookie consent for online news outlets. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–12.

Song, D., Lau, R. Y., Bruza, P. D., Wong, K.-F., & Chen, D.-Y. (2007). An intelligent information agent for document title classification and filtering in document-intensive domains. *Decision Support Systems*, *44*(1), 251–265.

Soumya George, K., & Joseph, S. (2014). Text classification by augmenting bag of words (bow) representation with co-occurrence feature. *IOSR Journal of Computer Engineering*, *16*(1), 34–38.

Stavrakakis, I., Curley, A., O'Sullivan, D., Gordon, D., & Tierney, B. (2021). A framework of web-based dark patterns that can be detected manually or automatically.

Sternberg, S. (2018). Why do courts craft vague decisions? evidence from a comparative study of court rulings in germany and france using quantitative text analysis. *University of Mannheim, Working Paper, available at https://sebastiansternberg. github. io/pdf/Sternberg_Value_of_Vagueness_CEL SE18. pdf*.

Sun, J., Tang, Z., Yin, H., Wang, W., Zhao, X., Zhao, S., Lei, X., Zou, W., & Li, X. (2021). Semantic data augmentation for end-to-end mandarin speech recognition. *arXiv preprint arXiv:2104.12521*.

Thavareesan, S., & Mahesan, S. (2020). Word embedding-based part of speech tagging in tamil texts. *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 478–482.

Tomas, F., Dodier, O., & Demarchi, S. (2022). Computational measures of deceptive language: Prospects and issues. *Frontiers in Communication*, *7*, 792378.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). (un)informed consent: Studying gdpr consent notices in the field. *Proceedings of the 2019 acm sigsac conference on computer and communications security*, 973–990.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

VM, N., Kumar R, D., et al. (2019). Implementation on text classification using bag of words model. *Proceedings of the Second International Conference on Emerging Trends in Science & Technologies For Engineering Systems (ICETSE-2019)*.

Volkova, S., & Jang, J. Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media. *Companion Proceedings of the The Web Conference 2018*, 575–583.

European Data Protection Board. (2020). *Guidelines 05/2020 on consent under regulation 2016/679* (tech. rep.) [https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf].

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-based bag-of-words model for text classification. *IEEE Access*, *8*, 82641–82652.

Yang, L., Li, C., Ding, Q., & Li, L. (2013). Combining lexical and semantic features for short text classification. *Procedia Computer Science*, *22*, 78–86.

Yao, Y., & Huang, Z. (2016). Bi-directional lstm recurrent neural network for chinese word segmentation. *International conference on neural information processing*, 345–353.

Ye, Z., Geng, Y., Chen, J., Chen, J., Xu, X., Zheng, S., Wang, F., Zhang, J., & Chen, H. (2020). Zero-shot text classification via reinforced self-training. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3014–3024.

Yin, W., Hay, J., & Roth, D. (2019a). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, *abs/1909.00161*. http://arxiv.org/abs/1909.00161

Yin, W., Hay, J., & Roth, D. (2019b). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Zagal, J. P., Björk, S., & Lewis, C. (2013). Dark patterns in the design of games. *Foundations of Digital Games 2013*.

Zuo, X., Chen, Y., Liu, K., & Zhao, J. (2019). Event co-reference resolution via a multi-loss neural network without using argument information. *Science China Information Sciences*, *62*(11), 1–9.