MIS Quarterly

RESEARCH ARTICLE

AUTOMATED ANALYSIS OF CHANGES IN PRIVACY POLICIES: A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING APPROACH[1]

Fangyu Lin
Department of Information Systems and Cyber Security, Alvarez College of Business, The University of Texas at San Antonio
San Antonio, TX, U.S.A {fangyu.lin@utsa.edu}

Sagar Samtani
Department of Operations and Decision Technologies, Kelley School of Business, Indiana University
Bloomington, IN, U.S.A {ssamtani@iu.edu}

Hongyi Zhu
Department of Information Systems and Cyber Security, Alvarez College of Business, The University of Texas at San Antonio
San Antonio, TX, U.S.A {hongyi.zhu@utsa.edu}

Laura Brandimarte and Hsinchun Chen
Department of Management Information Systems, Eller College of Management, The University of Arizona
Tucson, AZ U.S.A. {lbrandimarte@arizona.edu} {hsinchun@arizona.edu}

The increasing societal concern for consumer information privacy has led to the enforcement of privacy regulations worldwide. In an effort to adhere to privacy regulations such as the General Data Protection Regulation (GDPR), many companies' privacy policies have become increasingly lengthy and complex. In this study, we adopted the computational design science paradigm to design a novel privacy policy evolution analytics framework to help identify how companies change and present their privacy policies based on privacy regulations. The framework includes a self-attentive annotation system (SAAS) that automatically annotates paragraph-length segments in privacy policies to help stakeholders identify data practices of interest for further investigation. We rigorously evaluated SAAS against state-of-the-art machine learning (ML) and deep learning (DL)-based methods on a well-established privacy policy dataset, OPP-115. SAAS outperformed conventional ML and DL models in terms of F1-score by statistically significant margins. We demonstrate the proposed framework's practical utility with an in-depth case study of GDPR's impact on Amazon's privacy policies. The case study results indicate that Amazon's post-GDPR privacy policy potentially violates a fundamental principle of GDPR by causing consumers to exert more effort to find information about first-party data collection. Given the increasing importance of consumer information privacy, the proposed framework has important implications for regulators and companies. We discuss several design principles followed by the SAAS that can help guide future design science-based e-commerce, health, and privacy research.

Keywords: Privacy policy, structured self-attentive sentence embedding, deep learning, attention mechanisms, multi-label classification, GDPR, privacy analytics, computational design science

[1] Gedas Adomavicius was the accepting senior editor for this paper. Nan Zhang served as the associate editor.

Introduction

The rapid proliferation of e-commerce, social media, and other web services has enabled an unprecedented number of consumers to share large quantities of personal information on the internet. As a result, consumer information privacy has rapidly emerged as a significant societal issue (Acquisti et al., 2020). Increasing concern about how companies maintain the information privacy of their consumers has led to the development, update, and enforcement of privacy regulations such as the EU General Data Protection Regulation (GDPR) in 2018 and the California Consumer Privacy Act (CCPA) in 2020. Each regulation stipulates how companies must control customers' personal information. Companies that violate regulations can incur significant financial fines and damage their reputation. We summarize major companies that have recently violated GDPR in Table 1. Events are summarized based on the company name, industry type, country whose court decided to fine the company, fine incurred, and violation.

The far-reaching implications of regulations worldwide have led to many companies revising their privacy policies to include details about their data practices (i.e., collecting, processing, storing, sharing, and protecting customer data). Privacy policy revisions often result from the introduction of new regulations (e.g., GDPR, CCPA) or from requirements stipulated by the regulation. For example, the Federal Deposit Insurance Corporation (FDIC) and CCPA require banks and companies to update their privacy policies at least once per year (Bowers et al., 2017). Many regulations require companies to update their policies after introducing a new product or service. The frequency of updates can often cause the length of privacy policies to increase rapidly. Between 2009 and 2019, the average length of updated privacy policies doubled (Amos et al., 2021). In Figure 1, we present Amazon's privacy policy pre-GDPR (March 3, 2014) and post-GDPR (February 12, 2021) to illustrate how the privacy policy grew in length and complexity.

In the "For What Purposes Does Amazon Use Your Personal Information" section of Amazon's policy, the text related to using consumers' information for improving Amazon services in Amazon's pre-GDPR privacy policy contained only three words (Red Box 1 in Figure 1). However, following the implementation of GDPR, the number of words pertaining to the same purpose increased to 28 (Red Box 2 in Figure 1). In addition, the number of legalistic, jargon-laden, and ambiguous phrases increased (e.g., "comply with legal obligations") (Red Box 3 in Figure 1). These characteristics have caused legislators and researchers to become increasingly concerned that companies may draft privacy policies in ways that are compliant with regulations but do not actually improve consumers' ability to understand and control how companies process their personal information (Fazzini, 2019). Moreover, the update frequency and ever-growing length of privacy policies have also created challenges for companies wishing to ensure that their policies comply with regulations and for regulators aiming to monitor and enforce regulations. Taken together, these concerns underscore the significant need to identify how the contents of privacy policies change and are presented based on the stipulations of new or existing privacy regulations. While information systems (IS) scholars are uniquely positioned to analyze the text characteristics of privacy policies, prevailing IS privacy literature has focused on privacy concerns, risks, and controls from behavioral and economic perspectives. Methods within these paradigms are not designed to analyze the rich text within privacy policies. Therefore, a novel information technology (IT) artifact equipped with advanced machine learning (ML) and deep learning (DL) methods for text analytics is needed to identify how privacy policies change.

In this study, we adopted the computational design science paradigm to design a novel privacy policy evolution analytics framework that analyzes long and complex privacy policy texts to help identify how text content changed after regulations were implemented. At the heart of this framework stands a novel DL-based self-attentive annotation system (SAAS) that draws upon emerging DL-based structured self-attentive sentence embedding (SSASE) and attention mechanism techniques. SAAS aims to automatically annotate paragraph-length segments in privacy policies into one or more data practice categories to help relevant stakeholders, particularly companies and regulators, identify specific aspects of privacy policies for further investigation (in this study, identifying how privacy policies changed according to new regulations or updates to existing regulations) without reading large amounts of text. There are two novelties in SAAS's design. First, a row-wise attention (RWA) mechanism aims to identify the set of words and phrases (i.e., aspects) that helps SAAS assign the correct data practice category label(s) for a privacy policy segment. Second, a multi-label classifier learns sharing parameters that introduce strong regularization effects to make models less prone to overfitting. We rigorously evaluated SAAS against prevailing methods in automated privacy policy analysis literature and benchmark ML and DL algorithms with a series of experiments. We demonstrated the practical utility of our proposed framework with an in-depth case study on Amazon's pre- and post-GDPR privacy policies. Apart from offering academic and practical contributions to IS privacy analytics, our proposed SAAS follows several key design principles that can guide the design of future IT artifacts for e-commerce and health analytics applications.

The remainder of this paper is organized as follows. First, we review literature related to IS information privacy research and computational design science guidelines, privacy policy analysis, SSASE, and attention mechanisms. Second, we summarize key research gaps within the extant literature and pose research questions for the study. Third, we present our proposed privacy policy evolution analytics framework. Fourth, we present the results of our experiments and case study. Finally, we discuss this study's contributions to the IS knowledge base, summarize selected managerial implications, and present some promising directions for future research.

Table 1. Selected Recent Cases of GDPR Violations

Company	Industry type	Country	Fine (€)	Violation
Google	Computer software	France	150M	Insufficient legal basis for data processing
WhatsApp	Instant messaging	Ireland	225M	Insufficient fulfillment of information obligations
Amazon	E-commerce	Luxembourg	746M	Noncompliance with general data processing principles
H&M	Retail	Germany	35M	Insufficient legal basis for data processing
TIM	Telecommunications	Italy	27.8M	Insufficient legal basis for data processing

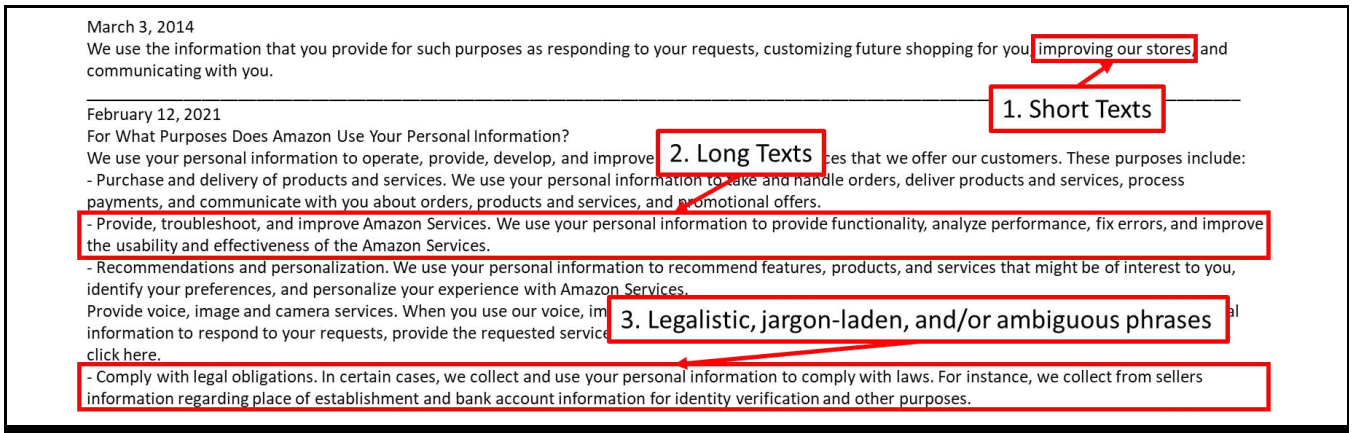

Figure 1. Amazon's Privacy Policy Pre- (Top) and Post- (Bottom) GDPR

Literature Review

We review four areas of literature to ground our research. First, we review recent IS information privacy research and the computational design science paradigm to guide the development of our proposed DL-based privacy policy evolution analytics framework. Second, we review privacy policy analysis literature to identify prevailing methods for automatically detecting changes to privacy policies. Third, we review SSASE to understand how a DL-based text analytics technique could be leveraged to enhance automated privacy policy evolution analytics. Finally, we identify attention mechanisms to identify approaches that can dynamically weigh input features within privacy policy text to improve SSASE performance.

IS Information Privacy Research and Computational Design Science Guidelines

Information privacy is fundamentally defined as "the ability of the individual to control personal information about one's self" (Stone et al., 1983). The explosive growth of personal information disclosure on the internet has motivated many IS scholars to scrutinize varying aspects of information privacy. To date, IS scholars have leveraged behavioral and economic paradigms to make remarkable progress in three major categories of information privacy research: concern, control, and risk. In Table 2, we summarize the focus, IS paradigm, and analytical method(s) leveraged in selected recent major IS information privacy studies.

Research examining privacy concerns, controls, and risks has primarily employed behavioral theories or econometric models to investigate the impact of privacy concerns on user or organizational decision-making, explore how various privacy controls influence users' behaviors, or examine the factors that affect privacy risk, respectively (Cao et al., 2018; Kim & Kwon, 2019; Wunderlich et al., 2019). Despite their important contributions, the analytical methods employed in these studies were not designed to analyze lengthy and unstructured privacy policy text. Consequently, a novel IT artifact designed to identify privacy policy evolution requires a principled approach. The design science paradigm offers prescriptive guidelines on designing, developing, and evaluating novel IT artifacts (e.g., constructs, models, methods, and instantiations) for critical societal applications (Hevner et al., 2004). Four genres of design science exist (Rai, 2017): computational, optimization, economics, and representation. Among the four, the computational genre is the most relevant for developing novel computational approaches, frameworks, models, and algorithms for advanced text analytics research.

Table 2. Summary of Recent Selected Major IS Information Privacy Literature

Category	Year	Author(s)	Focus	IS paradigm	Analytical method(s)*
Privacy concern	2021	Cichy et al.	The impact of privacy concern on connected car adoption	Behavioral	SEM
	2019	Wunderlich et al.	The impact of privacy concern on Internet of Things (IoT) adoption	Behavioral	Qualitative coding, hierarchical regression
	2019	Buckman et al.	Factors affecting users' valuation of their personal information	Behavioral	ANCOVA, Tobit regression
	2019	Crossler & Bélanger	Factors affecting the use of privacy settings on smartphones	Behavioral	SEM
	2018	Adjerid et al.	Examining rational cognition and heuristics of privacy decision-making	Behavioral	Linear regression
	2018	Gopal et al.	Users' privacy concerns about companies' third-party sharing strategies	Economic	Econometric model
	2017	Breward et al.	The impact of privacy and security concerns on controversial IT adoption	Behavioral	Qualitative coding, SEM
	2017	Koh et al.	The impact of privacy cost on voluntary profiling on e-commerce platforms	Economic	Econometric model
Privacy control	2021	Zalmanson et al.	The impact of social cues and trust on users' personal information disclosure	Behavioral	OLS and logistic regression path
	2018	Cao et al.	The impact of peer disclosure and related policies on online community participation	Economic	Econometric model
	2018	Gal-Or et al.	The impact of targeted ads and privacy controls on users' selection of platforms	Economic	Econometric model
	2018	Heimbach & Hinz	The impact of sharing mechanism on content sharing in social media	Behavioral	Logistic model, Poisson regression
	2016	Cavusoglu et al.	The impact of privacy control options on disclosure behavior on social media	Economic	Poisson regression
Privacy risk	2019	Kim & Kwon	The impact of electronic health records (EHRs) and meaningful use on the risk of patient information breaches	Economic	Cox proportional hazards model

Note: *ANCOVA: analysis of covariance; OLS: ordinary least squares; SEM: structural equation model

IT artifacts developed through the lens of the computational design science paradigm generally follow three guidelines (Rai, 2017). First, the artifact's design can be inspired by key domain requirements or characteristics when a strong underlying theory is lacking. For example, in Li and Qin (2017), unique data characteristics guided the development of a novel text analytics framework that incorporated carefully constructed feature representations and algorithms to anonymize medical records. Second, the artifact's novelty is demonstrated by evaluating its technical performance against state-of-the-art approaches via well-established quantitative metrics (e.g., precision, F1). Finally, the artifact should contribute to the IS knowledge base to help guide related future research. Contributions can include situated implementations (e.g., processes, software, etc.) and/or nascent design theory in the form of design principles. Properly executing each guideline requires a strong understanding of the application space for which the artifact is being developed. For this study, this requires reviewing key data characteristics of privacy policies and examining how advanced text analytics techniques can be developed to identify changes in privacy policies. Therefore, we review the extant privacy policy analysis literature next.

Privacy Policy Analysis

A privacy policy is a legal contract that describes a company's collecting, processing, sharing, and storing practices of users' personal information. It is currently the primary instrument stakeholders (e.g., regulators and companies) rely on to understand a company's data practices (Amos et al., 2021). In general, 10 major categories of data practices exist in a privacy policy (Wilson et al., 2016). We describe each data practice category and specify selected recent regulations requiring companies to disclose the category in their privacy policies in Table 3.

Table 3. Summary of Key Data Practice Categories in Privacy Policies

Category	Description	Regulations
First Party Collection (FP)	What data is collected and how and why a company collects data	GDPR, CCPA
Third Party Sharing/Collection (TP)	What data is shared and how and why a company shares data with third parties	GDPR, CCPA
User Choice/Control (UCC)	Privacy choices and control options available for users	GDPR, CCPA
User Access, Edit, Deletion (UAED)	How users can access, edit, or delete their data	GDPR, CCPA
Data Retention (DR)	How long user information is stored	GDPR, CCPA
Data Security (DS)	How user information is protected	GDPR, CCPA
Policy Change (PC)	How users will be informed about changes to the privacy policy	GDPR, CCPA
Do Not Track (DNT)	If and how Do Not Track signals for online tracking and advertising are honored	CCPA
International & Specific Audiences (ISA)	Practices that pertain only to one specific user group	HIPPA, COPPA
Other	Contact information, introduction, etc.	GDPR, CCPA

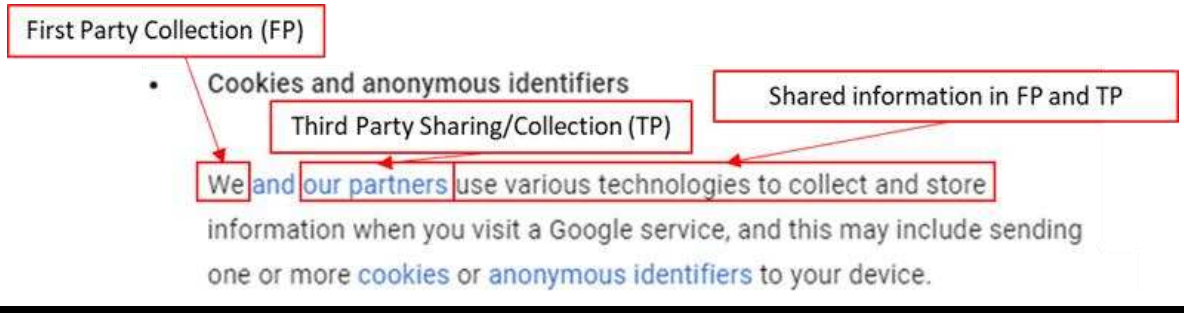

Figure 2. A Sample Segment in Google's Privacy Policy

First Party Collection (FP), Third Party Sharing/Collection (TP), Data Retention (DR), and Data Security (DS) detail what, how, and why first and third parties collect, process, store, share, and protect customer data. User Choice/Control (UCC), User Access, Edit, Deletion (UAED), and Do Not Track (DNT) pertain to a user's rights. International & specific audiences (ISA) are data practices that pertain only to a specific region or user group. A privacy policy often contains multiple segments (i.e., a set of consecutive and semantically coherent sentences) that present information about data practice categories (Wilson et al., 2016). Although recent privacy regulations clearly specify the requirements for information disclosure, there is no mandated or standard format that companies should follow when presenting their privacy policies (Alabduljabbar et al., 2021). As a result, companies often provide information for a specific data practice category in separate segments throughout their privacy policy. Moreover, companies may often use one segment to explain multiple data practice categories. We present a sample segment in Google's privacy policy that pertains to two categories in Figure 2.

The selected segment of Google's privacy policy contains details about FP (indicated by the word "We"), TP (indicated by the phrase "our partners"), and content about both FP and TP (indicated by the phrase "use various technologies to collect and store"). Dispersed and mixed information about data practices in segments can make privacy policy analysis a nontrivial task (Degeling et al., 2019; Linden et al., 2020). Furthermore, privacy regulations often require companies to disclose and regularly update each practice in their privacy policies to help users understand their rights to control their data. To comply with the regulations, segments in privacy policies often change, evolve, and grow in length (Amos et al., 2021). These changes can exacerbate the challenge for companies to manage and evaluate their compliance and regulators aiming to enforce regulations. We review selected recent privacy policy analysis research in Table 4 to understand what existing approaches have been proposed to help researchers, companies, and regulators analyze privacy policies. The summary is organized based on the focus of the study, the dataset examined, the corresponding privacy regulations, and the methodology employed.

Table 4. Summary of Selected Recent Privacy Policy Analysis Literature

Year	Author	Focus	Dataset			Regulation**	Methodology				
			Source*	# of policies	Time periods		Manual analysis	Automated analysis			
								Readability metrics	Keyword searching	Descriptive statistics	NLP***
2022	Arora et al.	Comparative study	English and German websites	155	1	GDPR, CCPA	No	No	No	Yes	BERT
2021	Amos et al.	Comparative study	English websites	1M	22	GDPR	No	Yes	Yes	Yes	No
2021	Qamar et al.	Compliance detection	OPP-115	115	1	GDPR, PDPA	No	No	No	No	BOW + LR, SVM, BERT
2021	Zaeem & Barber	Comparative study	OPP-115	115	1	No	No	No	No	No	CNN, NB
2020	Akanfe et al.	Privacy risk assessment	Mobile wallets and remittance apps	353	1	GDPR	No	No	Yes	No	No
2020	Akanfe et al.	Privacy risk assessment	Mobile wallets and remittance apps	230	1	GDPR	No	No	Yes	No	BOW, LDA
2020	Kumar et al.	Privacy settings assistant	English websites	6K	1	GDPR, CCPA	No	No	No	Yes	LR, topic modeling, BERT, TF-IDF
2020	Linden et al.	Comparative study	OPP-115	115	2	GDPR	Yes	Yes	Yes	Yes	CNN
2019	Andow et al.	Compliance detection	Apps from Google Play Store	11K	1	No	No	No	Yes	No	Parse tree + Rule-based
2019	Chang et al.	Privacy settings assistant	OPP-115	115	1	GDPR	No	No	No	No	CNN, RF
2019	Degeling et al.	Comparative study	EU websites	112K	12	GDPR	Yes	No	Yes	Yes	No
2019	Fawaz et al.	Comparative study, risk assessment, privacy settings assistant	OPP-115	115	2	GDPR	No	No	No	No	CNN
2019	Kumar et al.	Data practice annotation	OPP-115	115	1	No	No	No	No	No	FastText, LR, MLP, CNN, BERT
2019	Nejad et al.	Privacy risk assessment	OPP-115	115	1	GDPR	No	No	Yes	No	Did not specify
2019	Ravichander et al.	QA system	Apps from Google Play Store	35	1	No	No	No	No	No	SVM, CNN, BERT
2019	Story et al.	Compliance detection	Apps from Google Play Store	1M	1	No	No	No	No	Yes	BOW + Feature engineering + SVM
2019	Zimmeck et al.	Compliance detection	Apps from Google Play Store	1M	1	GDPR, COPPA, CalOPPA	No	No	No	Yes	BOW + SVM
2018	Gopinath et al.	Document segmentation	English websites	152	1	No	No	No	No	No	K-means, feature engineering, MLP
2018	Harkous et al.	QA system	OPP-115	115	1	No	No	No	No	No	CNN
2018	Tesfay et al.	Privacy risk assessment	EU websites	45	1	GDPR	No	No	No	No	BOW + NB, SVM, DT, RF

2018	Story et al.	Comparative study	Apps from Google Play Store	3M	3	CalOPPA, DOPPA, FIPPs	No	No	No	Yes	No
2017	Evans et al.	Privacy risk assessment	English websites	30	1	EU Directive 95/46/EC, HIPAA	No	No	No	No	Regular expressions, parse tree
2017	Nisal et al.	Privacy settings assistant	OPP-115	115	1	No	No	No	No	No	Feature engineering, LR
2017	Zaeem & Barber	Comparative study	NYSE, Nasdaq, and AMEX	600	1	FIPPs, COPPA	Yes	No	No	No	No
2018	Oltramari et al.	QA system	OPP-115	115	1	No	No	No	No	No	Rule-based
2017	Sathyendra et al.	Privacy settings assistant	OPP-115	115	1	FIPPs	No	No	Yes	No	BOW, LDA, Parse tree + LR
2016a	Bhatia et al.	Privacy risk assessment	English websites	5	1	No	No	No	No	No	Regular expressions, parse tree
2016b	Bhatia et al.	Privacy risk assessment	English websites	15	1	EU Directive 95/46/EC, HIPAA	No	No	No	No	Regular expressions, parse tree
2016	Liu et al.	Data practice annotation	OPP-115	115	1	No	No	No	No	No	NMF, BOW + LR, LDA
2016	Sathyendra et al.	Privacy settings assistant	OPP-115	115	1	FIPPs	No	No	Yes	No	BOW + LR, SVM, RF, NB, KNN
2016	Slavin et al.	Compliance detection	Apps from Google Play Store	477	1	FIPPs	Yes	No	Yes	Yes	No
2017	Zimmeck et al.	Compliance detection	Apps from Google Play Store	17K	1	CalOPPA, DOPPA, FIPPs, COPPA	No	No	Yes	Yes	BOW + LR, SVM

Note: *NYSE, Nasdaq, and AMEX: Stock Exchange Websites; OPP-115: Online Privacy Policies, set of 115. **CalOPPA: California Online Privacy Protection Act; COPPA: Children's Online Privacy Protection Act; DOPPA: Delaware Online Privacy and Protection Act; FIPPs: Federal Trade Commission's Fair Information Practice Principles; HIPAA: Health Insurance Portability and Accountability Act; PDPA: Personal Data Protection Act. ***BERT: bidirectional encoder representations from transformers; BOW: bag-of-words; CNN: convolutional neural network; DT: decision tree; KNN: *k*-nearest neighbors; LDA: latent Dirichlet allocation; LR: logistic regression; MLP: multi-layer perceptron; NB: naive Bayes; NMF: non-negative matrix factorization; RF: random forest; SVM: support vector machine; TF-IDF: term frequency-inverse document frequency.

Extant privacy policy analysis literature covers several major themes, including compliance detection, privacy risk assessment, privacy settings assistants, and comparative studies. Compliance detection studies have typically employed classical ML methods with bag-of-words representations to examine whether a company's data practices comply with privacy regulations (Qamar et al., 2021; Story et al., 2019). Privacy risk assessment studies have employed classical ML or keyword searching to evaluate overall user privacy risks based on the types and amount of personal information collected and third-party sharing (Akanfe et al., 2020b; Fawaz et al., 2019). Studies on privacy settings assistants have employed unsupervised topic modeling, classical ML methods, and parse trees to focus specifically on opt-out/opt-in options (Kumar et al., 2020; Sathyendra et al., 2016). Comparative studies compare privacy policies across different times (Amos et al., 2021), languages (Arora et al., 2022), and organizations (Zaeem & Barber, 2021). Since the focus of our research is on comparative studies, we discuss these studies in further detail.

Most past researchers executing comparative analysis studies have employed manual analyses (Zaeem & Barber, 2017), readability metrics, keyword searching, descriptive statistics (Story et al., 2018), or a combination thereof (Amos et al., 2021; Degeling et al., 2019). The most common dataset used in comparative studies is the "Online Privacy Policies, set of 115" (OPP-115) (Wilson et al., 2016). Developed by the Usable Privacy Policy Project at Carnegie Mellon University, OPP-115 includes 115 English privacy policies published between 2003 and 2015 from well-known, highly ranked websites across 15 sectors (as defined by DMOZ.org), as determined by Google trends. OPP-115 is suitable for comparative studies because the annotation scheme covers all 10 data practice categories and focuses on segments rather than individual sentences. These characteristics allow for a more thorough elaboration of all data practice categories and can facilitate a more comprehensive evaluation. However, privacy policies are lengthy and lack a standard format. Since prevailing methods for comparative analysis can result in incomplete content extraction and have limited scalability,

recent comparative analysis studies have employed supervised DL algorithms, namely CNN (Zaeem & Barber, 2021; Linden et al., 2020). DL-based supervised learning techniques have been shown to effectively learn from the well-defined data practice categories by privacy researchers and label privacy policy datasets to automatically identify changed/different data practice information about stakeholders' interest in privacy policies. Thus, we focus on supervised learning techniques.

Existing studies employing supervised learning techniques first manually or automatically segment privacy policies (Kumar et al., 2020; Harkous et al., 2018). Then, a data practice annotation system based on ML or DL algorithms (Harkous et al., 2018; Kumar et al., 2019; Liu et al., 2016) is often used to annotate segments. This process proceeds as follows:

Step 1: Segment privacy policies using segmentation tools. While sentence-level segmentation is suitable for identifying information type or opt-in/opt-out options (Kumar et al., 2020), paragraph-length segments are required to comprehensively elaborate all data practices (Wilson et al., 2016). Automated segmentation tools segment privacy policy text into paragraph-length segments based on HTML tags or ML (Harkous et al., 2018). HTML-tag-based segmentation may result in semantically incoherent segments. Thus, ML-based tools that merge adjacent sentences with high semantic similarities to generate coherent segments are preferred (e.g., GraphSeg (Glavas et al., 2016)).

Step 2: Annotate segments using ml algorithms. Following segmentation, conventional ML algorithms (e.g., LR, SVM) are then adopted to train 10 binary classifiers, each to predict (annotate) if a segment belongs to one or more (i.e., multi-label classification) of the 10 data practice categories (Wilson et al., 2016).

Step 3: Analyze annotated segments. Based on the annotated segment, stakeholders (e.g., companies and regulators) conduct targeted (downstream) analyses about the specific components of privacy policies (e.g., pre-post analyses, etc.).

Existing privacy policy analysis studies leveraging conventional ML algorithms (e.g., LR, SVM, NB, DT, RF; see Kumar et al., 2019; Qamar et al., 2021; Zaeem & Barber, 2021) for data practice annotation often suffer from low annotation accuracy due to their reliance on segment representations generated by bag-of-words (BOW) as input (Sathyendra et al., 2016, 2017). Such representations assume that segments in the same data practice category share similar word distributions. In reality, segments often have diverse word choices. Consequently, BOW representations may lead to incorrect predictions due, in particular, to missing one or more data practice categories (labels) for a segment or misclassifying a segment into the wrong category. The former problem may lead regulators to conclude that a company has failed to address a regulated data practice in its policy, thus identifying a violation and issuing unwarranted fines. The latter issue can increase the effort needed to review misclassified segments or lead to a misunderstanding of companies' data practices (e.g., misinterpreting data practices as TP instead of FP).

Scholars have started adopting DL-based methods (e.g., multi-layer perceptron (MLP), CNN, BERT) for data practice annotation (Chang et al., 2019; Harkous et al., 2018; Linden et al., 2020; Zaeem & Barber, 2021). DL-based methods apply multiple layers of nonlinear transformations to automatically learn features from input text represented by word embeddings. The MLP and averaged word embeddings method (Kumar et al., 2019), while effective in many tasks, can compromise sequence information, neglect long-term dependencies, and struggle to capture complex sentence structures and the contextual nuances of each word within a segment. While CNNs have attained superior performance over conventional ML algorithms, these methods can often miss capturing long-range sequential dependencies (Yin et al., 2017). Such methods could still misclassify long data practice segments. BERT (Arora et al., 2022; Kumar et al., 2019; Qamar et al., 2021), despite its power, requires substantial fine-tuning data, which poses challenges when certain data practice categories, such as data retention, have limited instances in the dataset (OPP-115). These limitations necessitate an alternative DL-based approach that can capture long-range semantics dependencies to generate better segment representation for data practice segment annotation with limited instances. In recent years, RNN-based approaches such as long short-term memory (LSTM) have been extensively used to capture long-range sequential dependencies (Yin et al., 2017). Increasingly, scholars are improving the performance of RNN-based approaches by capturing the nonsequential global dependencies of input features with attention mechanisms (Lin et al., 2017). SSASE, which incorporates RNN-based processing with an attention mechanism, is a possible, suitable, and high-performing approach for processing data practice segments with lengthy and mixed information.

Structured Self-Attentive Sentence Embedding (SSASE)

SSASE generates text representations based on a bidirectional LSTM (BiLSTM) model with a multi-head self-attention mechanism (Lin et al., 2017). BiLSTM is a high-performing DL model often employed in text analytics tasks to capture sequential and contextual dependency information from text input. BiLSTMs have consistently outperformed CNN-based

methods in text analytics tasks where input texts may have long-range dependencies or when the prediction is based on the semantics of the entire text input (Yin et al., 2017). The multi-head self-attention mechanism extracts the nonsequential global dependencies of the inputs that the conventional BiLSTM may not capture. We depict the key SSASE operations for one data practice category in Figure 3.

SSASE's input data practice segment has n words, represented by a sequence of word embeddings $\mathbf{S} = (\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_n})$. BiLSTM processes the word embedding sequence in both forward and backward directions. Each direction generates a u-dimension "directional" hidden state $\overrightarrow{\mathbf{h_i}}$ (or $\overleftarrow{\mathbf{h_i}}$) based on $\mathbf{w_i}$ and the previous hidden state $\overrightarrow{\mathbf{h_{i-1}}}$ (or $\overleftarrow{\mathbf{h_{i+1}}}$). By concatenating $\overrightarrow{\mathbf{h_i}}$ and $\overleftarrow{\mathbf{h_i}}$ from the forward and backward directions, $\mathbf{h_i} = \left[\overrightarrow{\mathbf{h_i}}\, \overleftarrow{\mathbf{h_i}}\right]^T$ captures a more comprehensive summary of the current hidden state than using a single direction alone. All hidden states are denoted by a matrix $\mathbf{H} = (\mathbf{h_1}, \mathbf{h_2}, \ldots \mathbf{h_n})$, with a size of $n \times 2u$.

The multi-head self-attention mechanism learns weights by projecting $\mathbf{H}$ to different vector subspaces that each focus on a distinct aspect of the input segment. It generates a multi-head self-attention weight matrix $\mathbf{A}$ with a size of $r \times n$ by $\mathrm{Softmax}(\mathbf{W_{s2}} \tanh(\mathbf{W_{s1}} \mathbf{H^T}))$. $\mathbf{H^T}$ is a transposed hidden state matrix. $\mathbf{W_{s1}}$ and $\mathbf{W_{s2}}$ are two trainable weight matrices with shapes of $d_a \times 2u$ and $r \times d_a$, respectively, where d_a is a predefined hyperparameter and r is the number of attention heads. $\mathbf{A}$ summarizes "how much attention" should be paid to each word according to different aspects of the segment learned by attention heads. Each column in $\mathbf{A}$ corresponds to one word, while each row (i.e., head) is expected to highlight a salient set (i.e., aspect, component) of related words or phrases in the segment. A penalty term $\mathbf{P}$ in the loss function diversifies attention heads to avoid learning duplicate aspects: $\mathbf{P} = \|\mathbf{A}\mathbf{A^T} - \mathbf{I}\|_F^2$, where $\mathbf{I}$ is the identity matrix and $\|\cdot\|_F$ is the Frobenius norm. The penalty term is jointly minimized with the loss function for classification. Segment embedding $\mathbf{M}$ ($r \times 2u$) is the matrix product of $\mathbf{A}$ and $\mathbf{H}$. Each row in $\mathbf{M}$ encodes an aspect learned by the corresponding attention head. $\mathbf{M}$ is flattened into a vector and fed into a fully connected (FC) layer and a softmax layer for binary classification.

SSASE and its variants have outperformed conventional ML-based and DL-based algorithms in multi-class text classification applications, including categorizing health records (Chen et al., 2022) and analyzing sentiment in social media posts (Alagha, 2022). However, little work has examined SSASE's performance for multi-label classification tasks like data practice segment annotation. SSASE could potentially be formulated for multi-label classification tasks by training a separate binary classification model for each label with the sigmoid activation function (extensively used for generating binary predictions) being used in the fully connected output layer (Krohn et al., 2020). Such formulation is consistent with the concept of binary relevance (BR) in multi-label classification. BR operates by decomposing the multi-label classification process into a set of independent binary classification tasks (a classifier per label) (Zhang et al., 2018). BR operates well when minimal correlations between labels exist. However, data practice categories often possess interdependencies wherein all categories are related to the same topic of privacy. Previous studies have shown that multi-task learning can benefit loosely related text classification tasks (Harkous et al., 2018; Kerinec et al., 2018), which may represent one potential solution to overcoming the limitation of BR formation. When multiple tasks are closely related, multi-task learning learns sharing parameters that introduce strong regularization effects, making models less prone to overfitting each specific data practice category compared to independent binary classifiers.

In addition to the limitations associated with BR, SSASE treats all aspects of segments learned by the multi-head self-attention mechanism equally when generating segment embedding. However, some aspects are more differentiating than others in a data practice segment. For example, "share" indicates TP, while "the types of collected personal information" can indicate both FP and TP. Capturing these differentiating aspects could improve data practice segment annotation. An attention mechanism is a promising approach that can learn weights for different inputs based on their contributions to the model's final output (Vaswani et al., 2017). Therefore, we review attention mechanisms next to understand how to adapt SSASE.

Attention Mechanisms

Attention mechanisms operate by mapping a query vector $\mathbf{q}$ and a set of key vector-value vector pairs $(\mathbf{k}, \mathbf{v})$ to an output vector $\mathbf{o}$ (Vaswani et al., 2017). $\mathbf{q}$ can be considered a representation of interest (e.g., output embedding), and $\mathbf{k}$ represents the input features. $\mathbf{o}$ is computed as a weighted $\mathbf{v}$, where weights are alignment scores calculated based on the relationship (e.g., similarity) between $\mathbf{q}$ and $\mathbf{k}$. Higher alignment scores are assigned if elements in $\mathbf{k}$ are closely related to $\mathbf{q}$. Evaluating an attention mechanism is typically executed by comparing the performance of a model with the proposed attention mechanism against the model without it on a ground-truth dataset (Galassi et al., 2021; Spliethöver et al., 2019).

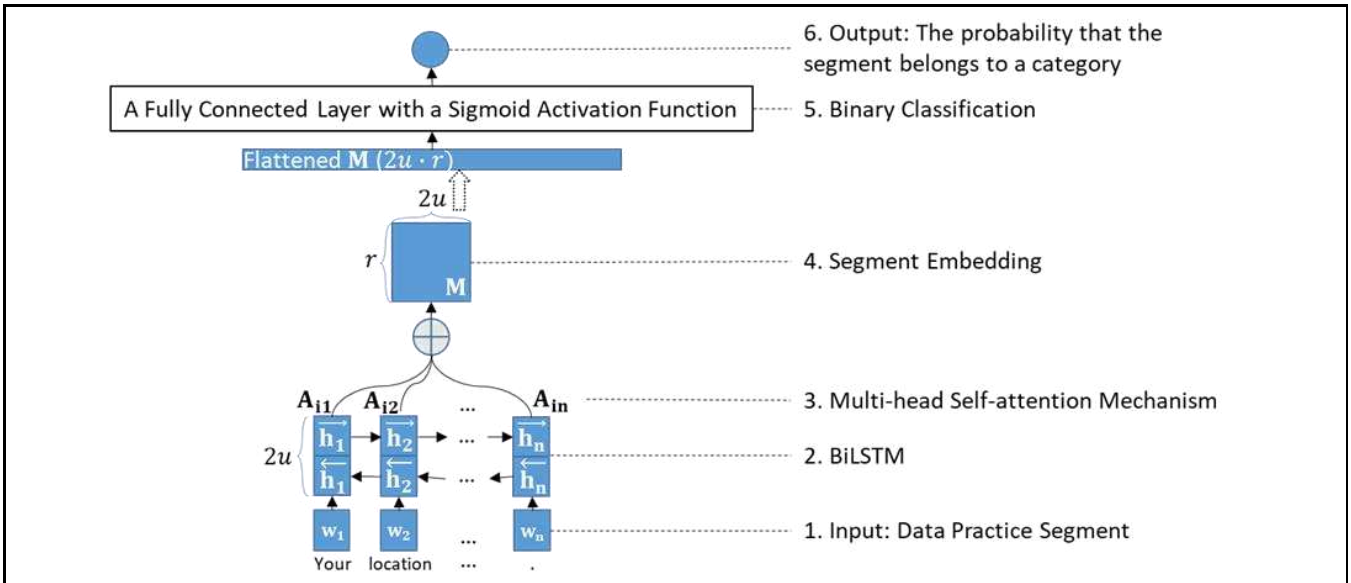

Figure 3. Conceptual Schematic of SSASE for One Data Practice Category* (Adapted from Lin et al., 2017)

Note: *The same procedure applies to each of the 10 data practice categories.

Attention mechanisms can be categorized into two major groups: general attention and self-attention (Vaswani et al., 2017). The former calculates the alignment score between $\mathbf{q}$ and $\mathbf{k}$, and the latter calculates the alignment score within the elements in K (i.e., $\mathbf{q} = \mathbf{k}$). Self-attention mechanisms have been incorporated into sequence models (e.g., BiLSTMs) to capture global feature dependencies for generating high-quality text representations in neural machine translation and sentiment analysis tasks (Letarte et al., 2018; Vaswani et al., 2017). The global feature dependencies are captured by relating input features at different positions of a sequence. Self-attention mechanisms dynamically weight input text features and can help identify the various informative aspects of data practice segments and produce an improved representation. However, how to incorporate self-attention into SSASE to improve generated segment representation and data practice segment annotation requires further investigation.

Research Gaps and Questions

We identified several research gaps in the literature review. First, while IS scholars have made significant progress in multiple areas of information privacy research, methods adopted in prior IS literature were not designed to operate on privacy policies' rich and complex text. Since privacy policies are the main instruments that companies use to convey their data practices, there is a need for an automated approach to annotate segments in privacy policies (i.e., label portions of privacy policies into their data practice categories) such that relevant stakeholders (e.g., companies, regulators) can assess the impact of regulations on companies' privacy policies in a targeted fashion. However, many past privacy policy analysis studies employed conventional ML approaches that relied on BOW-based segment representations, which cannot capture the interdependencies or other important features within the text. While DL-based methods can automatically extract salient features from text data, extant studies have primarily leveraged approaches that often missed long sequential word dependencies in privacy policies. SSASE is a potential high-performing text analytics approach that can capture long sequential word dependencies and nonsequential global semantics dependencies. However, SSASE's multi-head self-attention mechanism was mainly adopted for multi-class classification tasks and may miss key differentiating aspects within data practice segments. Formulating SSASE's multi-head self-attention mechanism to capture differentiating aspects within a multi-label classification approach (needed for privacy policy segment annotation) requires further investigation. Based on these research gaps, we pose the following research questions for the study:

RQ1: *How can the SSASE's multi-head self-attention be enhanced to identify key differentiating aspects in data practice segments to improve the performance of multi-label data practice segment annotation?*

RQ2: *How can the enhanced automated data practice segment annotation system help analyze how privacy policies evolve (e.g., are revised) following the enforcement of a privacy regulation?*

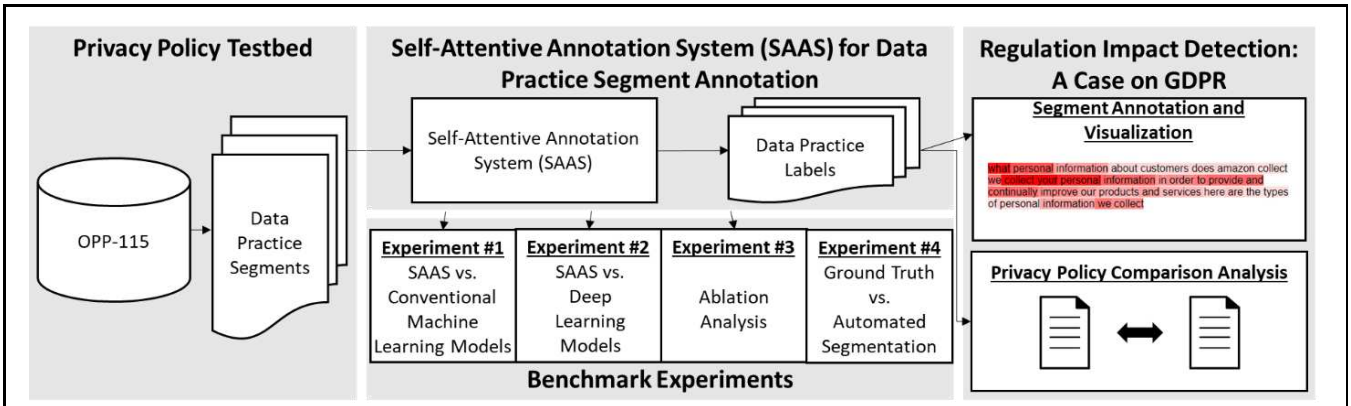

Figure 4. Proposed Privacy Policy Evolution Analytics Framework

Proposed Privacy Policy Evolution Analytics Framework

We propose a novel DL-based privacy policy evolution analytics framework to address the posed research questions. The proposed framework consists of four components (Figure 4): (1) Privacy Policy Testbed, (2) Self-Attentive Annotation System (SAAS) for Data Practice Segment Annotation, (3) Benchmark Experiments, and (4) Regulation Impact Detection: A Case on GDPR. We describe each component of the framework in the following subsections.

Privacy Policy Testbed

We adopted OPP-115 as our testbed. OPP-115 contains 115 English privacy policies with 3,792 segments from well-known websites. Each segment was annotated with one or more data practice labels from 10 categories by three law school students (Wilson et al., 2016). Although privacy policies in OPP-115 were collected before the release of several recent privacy regulations, the OPP-115 annotation scheme is consistent with the legal assumptions (e.g., transparency requirements, data practice categories, etc.) in recent regulations (e.g., GDPR, CCPA) (Leone & Di Caro, 2020; Poplavska et al., 2020). Consequently, OPP-115 is the prevailing dataset researchers have used for privacy policy analysis tasks, including training ML-based data practice segment annotation systems designed to annotate unlabeled privacy policies after regulations have been published (e.g., GDPR and CCPA) (Linden et al., 2020). Consistent with previous studies, we retained data practice category labels for each segment when two or more annotators agreed on labels (Harkous et al., 2018). As a result, 3,749 data practice segments that contained at least one data practice category label agreed upon by two or more annotators were included in our testbed.

Self-Attentive Annotation System (SAAS) for Data Practice Segment Annotation

Recognizing the key limitations of SSASE as it pertains to data practice segment annotation, we propose a novel SAAS that builds upon SSASE and attention mechanism principles. SAAS comprises a novel row-wise self-attentive embedding (RSE) model and a multi-label classifier. RSE extends SSASE with a novel attention mechanism to weigh the importance of different aspects in each row of segment embeddings learned by the multi-head self-attention mechanism. The multi-label classifier classifies RSE's learned representations into one or more data practice categories to help facilitate privacy policy evolution analysis. We present a comparison between the conventional SSASE and the proposed SAAS in Figure 5. The novelties of our proposed SAAS are highlighted in red. We summarize each SAAS component after.

Row-Wise Self-Attentive Embedding (RSE)

RSE takes a data practice segment as input. Each word in the segment is represented by a word embedding. Bi-directional gated recurrent units (BiGRU) is a variant of BiLSTM that automatically extracts the forward and backward context information from the embedding sequence. BiGRU achieves comparable performance while converging faster than BiLSTM, as it leverages a two-gate structure (as opposed to BiLSTM's three-gate structure) (Cho et al., 2014). The multi-head self-attention mechanism learns from all hidden states extracted by BiGRU the contributions of sets of related words or phrases to different aspects of the segment as the multi-head self-attention weight matrix. The multi-head self-attention weight matrix is applied back to the hidden states to produce the segment embedding **M**.

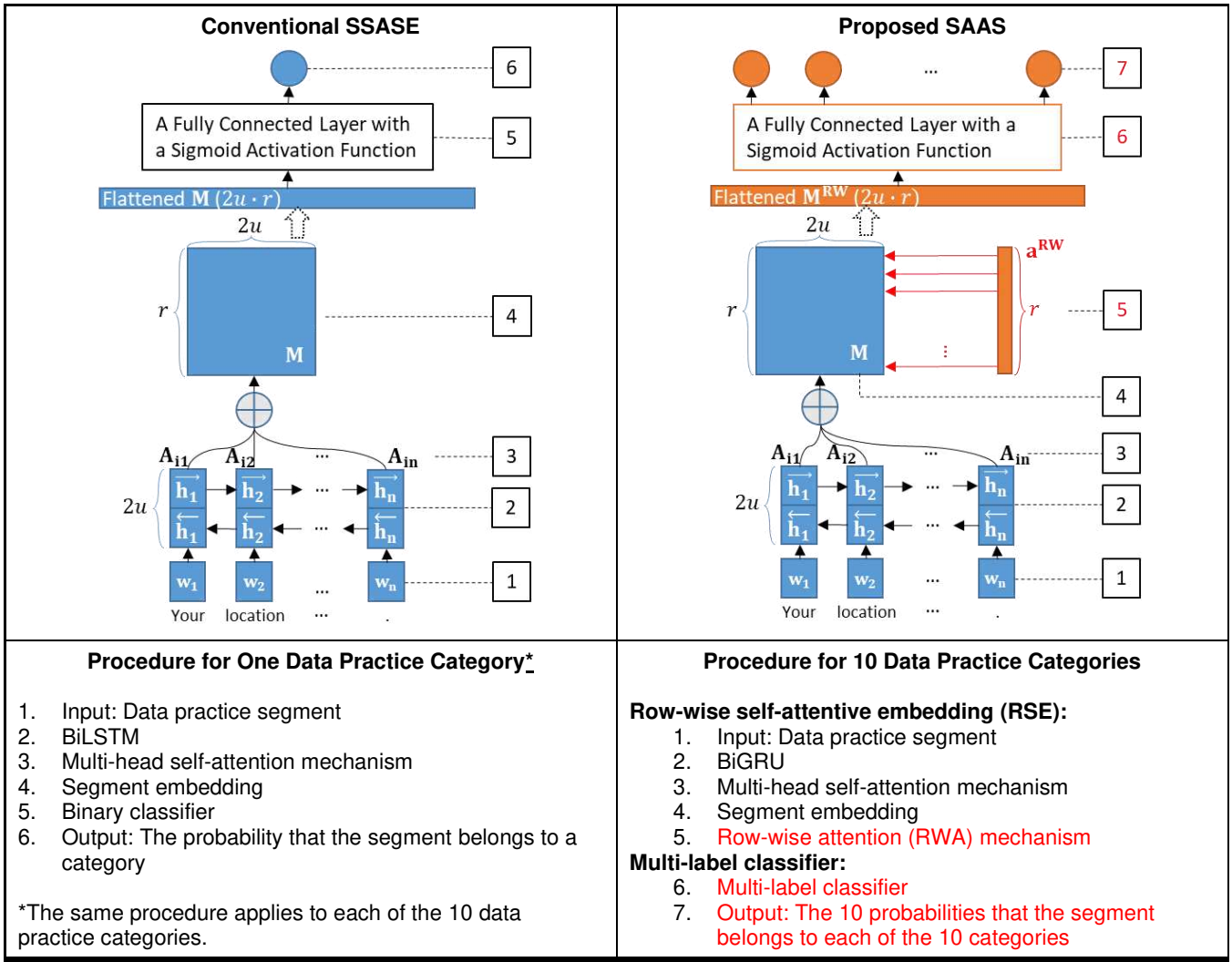

Figure 5. Illustration of Conventional SSASE and Proposed SAAS

While $\mathbf{M}$ is a low-dimensional matrix where each row encodes different aspects of the segment separately, not all aspects are equally important for predicting data practice labels. Therefore, we designed a novel row-wise attention (RWA) mechanism. Similar to how self-attention mechanisms determine each word's contribution to the model performance, the proposed RWA mechanism learns the importance of each row (aspect) in segment embedding that contributes to classification decisions as follows:

$$\mathbf{a}^{\mathbf{RW}} = \text{Softmax}(\mathbf{w}^{\mathbf{RW}}\mathbf{M}^{\top}),$$

where $\mathbf{a}^{\mathbf{RW}}$ is the row-wise attention weight vector, $\mathbf{M}^{\top}$ is the transposed segment embedding, and $\mathbf{w}^{\mathbf{RW}}$ is the trainable weight vector. The softmax activation function introduces nonlinearity into the row-wise attention weight vector. Each element in $\mathbf{a}^{\mathbf{RW}}$ indicates the importance of each aspect. All elements add up to 1. $\mathbf{a}^{\mathbf{RW}}$ is applied back to the segment embedding by multiplying each row in M with the corresponding weight in $\mathbf{a}^{\mathbf{RW}}$ formulating as follows:

$$\mathbf{M}^{\mathbf{RW}} = \text{Diag}(\mathbf{a}^{\mathbf{RW}})\mathbf{M},$$

where $\mathbf{M}^{\mathbf{RW}}$ is the weighted $\mathbf{M}$ with a size of $\mathrm{r} \times 2\mathrm{u}$ and $\text{Diag}(\mathbf{a}^{\mathbf{RW}})$ is a square diagonal matrix with the elements of vector $\mathbf{a}^{\mathbf{RW}}$ on the main diagonal. $\mathbf{a}^{\mathbf{RW}}$ is a regularizer in label predictions, i.e., if one aspect (row) is useful for label predictions, then all elements of such an aspect (row) are useful and should be assigned higher weights. RWA is suitable for complex text with varied lengths, such as data practice segments. $\mathbf{M}^{\mathbf{RW}}$ is then flattened into a vector and fed into the downstream multi-label classifier.

Table 5. Summary of Benchmark Experiments

Experiment	Model category*	Benchmark models*	References
SAAS vs. conventional machine learning models	Paragraph vector-based (Doc2Vec**): Sentence embedding is learned with a Continuous Bag of Words (CBOW) approach that predicts a missing word based on other words in the sentence.	Doc2Vec + LR	Wilson et al., 2016
		Doc2Vec + SVM	
		Doc2Vec + RF	Sathyendra et al., 2016
		Doc2Vec + KNN	
	Term frequency (TF)-based (TF-IDF): Sentence embedding is learned based on TF divided by inverse document frequency (IDF).	TF-IDF + LR	Sathyendra et al., 2016; Tesfay et al., 2018b; Zimmeck et al., 2019
		TF-IDF + SVM	
		TF-IDF + RF	
		TF-IDF + NB	
		TF-IDF + KNN	
SAAS vs. deep learning models	CNN-based: Text features are extracted through one convolutional layer and max pooling. Features are input into two dense layers for classification.	CNN	Harkous et al., 2018
	Uni-directional RNN-based: Contextual information is captured through a uni-directional recurrent structure. Extracted features are pooled through max and mean operations.	LSTM + Max pooling	Lai et al., 2015
		LSTM + Mean pooling	
		GRU + Max pooling	
		GRU + Mean pooling	
	Bi-directional RNN-based: Contextual information is captured through a bi-directional recurrent structure. Extracted features are pooled through max and mean operations.	BiLSTM + Max pooling	Lin et al., 2017
		BiLSTM + Mean pooling	
		BiGRU + Max pooling	
		BiGRU + Mean pooling	
	Attention-based: A self-attention mechanism and bi-directional recurrent structure learn a matrix embedding.	10 SSASEs, one for each data practice category	
		SSASE with a multi-label classifier	
Ablation analysis	Without RWA		Galassi et al., 2021; Spliethöver et al., 2019
	Replacing RWA with MLP		
	10 SAASs, one for each data practice category		
Ground truth vs. automated segmentation	Ground truth segmentation		Joty et al., 2013
	Automated segmentation		

Note: *Bi: bidirectional; CNN: convolutional neural network; GRU: gated recurrent unit; KNN: *k*-nearest neighbors; LR: logistic regression; LSTM: long-short term memory; MLP: multi-layer perceptron; NB: naive Bayes; RF: random forest; SVM: support vector machine. **Naive Bayes does not work with Doc2Vec text representation as naive Bayes requires positive numeric input.

Multi-Label Classifier

Originally, SSASE was only tested on various multi-class classification tasks. Although we could formulate a multi-label classification task into multiple binary classification tasks and train separate SSASE for each task, such a formulation would ignore relationships between categories. Therefore, we propose a multi-label classifier that incorporates 10 binary classification tasks in one model. The flattened M^{RW} is input into a fully connected layer with 10 output neurons. A sigmoid activation function, $\sigma(x) = \frac{1}{1+e^{-x}}$, transforms the outputs such that the outputs become 10 probabilities for 10 label predictions, $\hat{Y} = (\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_{10}})$, where $\widehat{y_i}$ is between 0 and 1 (how likely a segment belongs to the i-th data practice category). Binary cross entropy (BCE) is used as the loss function to compare each of the predicted probabilities to the actual label, which can be either 0 (i.e., does not belong to the category) or 1 (i.e., belongs to the category). BCE has been extensively used to evaluate binary classification tasks (Bird et al., 2009). BCE is formulated as $\mathrm{L} = -\frac{1}{N*10}\sum_{j=1}^{N}\sum_{i=1}^{10}\left(y_{ij}\log\left(\widehat{y_{ij}}\right) + \left(1-y_{ij}\right)\log\left(1-\widehat{y_{ij}}\right)\right)$, where N is the number of training instances, and y_{ij} is the ground-truth label of the i^{th} data practice category of the j^{th} training instance. By simultaneously minimizing the loss of 10 categories during backpropagation in model training, the model can learn common features across categories and is less prone to overfitting a specific category (Kerinec et al., 2018).

Benchmark Experiments

Consistent with the computational design science paradigm (Rai, 2017), we rigorously evaluated SAAS with four benchmark experiments (Table 5): (1) SAAS vs. conventional machine learning models, (2) SAAS vs. deep learning models, (3) ablation analysis, and (4) ground truth vs. automated segmentation.

Table 6. Number of Segments in Each Data Practice Category in OPP-115	
Data practice category (label)	**Number of Segments**
First Party collection (FP)	1,522
Third Party Sharing/Collection (TP)	1,186
User Choice/Control (UCC)	632
User Access, Edit, Deletion (UAED)	231
Data Retention (DR)	156
Data Security (DS)	375
Policy Change (PC)	192
Do Not Track (DNT)	32
International & Specific Audiences (ISA)	353
Other (O)	1,763
Total:	6,442

In Experiment 1, we compared SAAS against five conventional ML benchmark models commonly used in IS literature (Kitchens et al., 2018): logistic regression (LR), support vector machine (SVM), random forest (RF), naive Bayes (NB), and *k*-nearest neighbors (KNN). Two text representation techniques commonly used in privacy policy analysis literature, Doc2Vec and TF-IDF, represented each segment's text and were inputted into each model (Sathyendra et al., 2016; Wilson et al., 2016). In Experiment 2, we compared SAAS against four categories of DL-based models: CNN-based, uni-directional RNN-based (LSTM and GRU), bi-directional RNN-based (BiLSTM and BiGRU), and attention-based (SSASE). We considered two SSASE variants for data practice segment annotation: training 10 SSASEs (one for each data practice category) and one SSASE combined with a multi-label classifier. In Experiment 3, we conducted an ablation analysis that evaluated three variations of the SAAS: (1) SAAS without RWA, (2) SAAS replacing RWA with MLP, and (3) 10 SAAS models, one for each data practice category. In Experiment 4, we compared SAAS's performance on the original OPP-115 segmentation and ST-Ro segments to evaluate the effect of the automated segmentation on data practice annotation (Joty et al., 2013).

Consistent with privacy policy analysis literature, we executed each benchmark experiment with the 3,749 segments from the OPP-115 dataset that possessed one or more data practice labels agreed upon by two or more of the original annotators (Wilson et al., 2016). Overall, 2,848 segments had one label, 792 segments had two labels, 88 segments had three labels, 18 segments had four labels, and three segments had five labels. We present the number of segments associated with each data practice category (label) in Table 6.

The number of segments in each category ranged from 32 to 1,763, suggesting that the distribution of segments across the categories was imbalanced. The FP and TP categories had the highest number of segments, with 1,522 and 1,186, respectively, while the DNT category contained the fewest segments (32). We executed each experiment for each data practice category. Since the dataset was imbalanced, model performances were measured using precision, recall, and F1-score. Precision measures whether a model correctly classifies a segment into a specific category and is defined as follows:

$$Precision^{c_i} = \frac{TP(c_i)}{TP(c_i) + FP(c_i)},$$

where $c_i \in C$. C is the set of 10 data practice categories, $TP(c_i)$ (true positives) denotes the number of segments correctly classified to a specific data practice category c_i, and $FP(c_i)$ (false positives) is the number of segments incorrectly classified to a specific data practice category c_i. Recall measures whether a model detects all the segments in each data practice category and is defined as follows:

$$Recall^{c_i} = \frac{TP(c_i)}{TP(c_i) + FN(c_i)},$$

where $FN(c_i)$ (false negative) is the number of segments incorrectly classified as not a specific data practice category c_i. The F1-score is the harmonic mean of precision and recall and is defined as follows:

$$F1\text{-}score^{c_i} = \frac{2 \times Precision^{c_i} \times Recall^{c_i}}{Precision^{c_i} + Recall^{c_i}}.$$

We also used the micro-averaged precision, micro-averaged recall, and micro-averaged F1-score metrics to evaluate the performance of each model across all data practice categories (Harkous et al., 2018; Wilson et al., 2016). The three metrics are defined as follows:

$$Precision^{micro} = \frac{\sum_{c_i \in C} TP(c_i)}{\sum_{c_i \in C} TP(c_i) + FP(c_i)},$$

$$Recall^{micro} = \frac{\sum_{c_i \in C} TP(c_i)}{\sum_{c_i \in C} TP(c_i) + FN(c_i)}$$

$$F1\text{-}score^{micro} = \frac{2 \times Precision^{micro} \times Recall^{micro}}{Precision^{micro} + Recall^{micro}}.$$

Finally, we used hamming loss (HL) and micro-averaged HL to evaluate model performance for each category and across all categories. HL is a commonly used metric for evaluating the performance of multi-label classification tasks (Tsoumakas & Katakis, 2007). HL measures the fraction of labels that are incorrectly predicted. HL and micro-averaged HL are defined as follows:

$$Hamming\ Loss^{c_i} = \frac{FP(c_i) + FN(c_i)}{TP(c_i) + FP(c_i) + TN(c_i) + FN(c_i)}$$

$$Hamming\ Loss^{micro} = \frac{\sum_{c_i \in C} FP(c_i) + FN(c_i)}{\sum_{c_i \in C} TP(c_i) + FP(c_i) + TN(c_i) + FN(c_i)},$$

where $TN(c_i)$ (true negatives) denotes the number of segments correctly classified as not belonging to a specific data practice category c_i. HL examines how likely the model is to predict data practice segments with incorrect data practice categories.

5 times 2-fold (5×2) cross-validation was adopted for each experiment as it is suitable for comparing two classifiers on a single dataset (Demšar, 2006; Dietterich, 1998). 5×2 cross-validation overcomes the problem of underestimated variance and elevated Type I error when using resampled paired *t*-test and the *k*-fold cross-validated paired *t*-test. We randomly assigned all 3,749 segments into two partitions with an iterative-stratification sampling strategy (Sechidis et al., 2011) to ensure low performance variance across folds. In each fold, one partition was used for testing, and the other was used for training (90%) and validation (10%). This process was repeated five times, and the results were averaged to produce a single estimation. Paired *t*-tests were used to identify statistically significant differences between performance metrics (Demšar, 2006; Dietterich, 1998). Performance differences were considered significant at $p < 0.05$, 0.01, and 0.001.

All experiments were executed on a Microsoft Windows 10 Pro server with 128GB of random access memory, an Nvidia GeForce GTX 1070 Ti graphical processing unit, and an E5-2670 v4 at 2.60 gigahertz Intel central processing unit. All implementations were based on the PyTorch (Paszke et al., 2019), Natural Language Toolkit (Bird et al., 2009), NumPy (van der Walt et al., 2011), Pandas (McKinney, 2010), genism (Rehurek & Sojka, 2010), and scikit-learn (Pedregosa et al., 2011) packages. The complete details for SAAS appear in Appendix A.

Results and Discussion: Experiments and Case Study

Experiment 1 Results: SAAS vs. Conventional Machine Learning Models

Experiment 1 compared the performance of SAAS against conventional ML models. We present each model's performance in terms of micro-averaged precision, micro-averaged recall, micro-averaged F1-score, and micro-averaged HL in Table 7. Results are grouped based on the underlying text representation used by the algorithm. The best scores are highlighted in boldface. Each model's performance in each of the 10 data practice categories is reported in Appendix B.

SAAS outperformed all benchmark methods in terms of micro-averaged recall (0.714), micro-averaged F1-score (0.758), and micro-averaged HL (0.058) by statistically significant margins. Similarly, SAAS outperformed all methods, except TF-IDF + RF, on micro-averaged precision (0.807) by statistically significant margins. Overall, models generating segment representations that retained word-level information (i.e., term frequency-based and proposed SAAS) outperformed those that aggregated word-level information (i.e., paragraph vector-based) across all metrics. In privacy policy annotation, keywords are useful for identifying whether a segment belongs to a specific data practice category. For example, the keyword "share" is more likely to indicate "Third Party Sharing/Collection," and the keyword "collect" can indicate both "First Party Collection" and "Third Party Sharing/Collection." Within the term frequency-based category, LR and SVM outperformed other models on micro-averaged F1-score, attaining scores of 0.721 and 0.700, respectively, possibly due to their ability to process high-dimensional features (Kamath et al., 2018).

SAAS's performance is likely attributable to its ability to capture keywords, leverage each word's contextual and local semantics, and process high-dimensional feature sets. We present a segment SAAS correctly labeled as DS and FP, but TF-IDF + LR (the best-performing benchmark model in terms of F1-Score) mislabeled as FP only in Table 8. Instances related to FP were selected because they have the highest number of labels (653). The table is organized based on the company that the privacy policy belongs to, the segment in the privacy policy, the row-wise attention weight, the ground truth data practice category of the segment, and the predicted data practice category produced by SAAS and TF-IDF + LR. The color shades represent the normalized word weights learned by the attention head that extracted the aspect of the segment with the highest row-wise attention weight. Dark red indicates the higher importance of phrases in the segment. For illustration purposes, we set the number of attention heads as 5. If aspects are treated equally, each will have a row-wise attention weight of 0.2.

Table 7. Experiment 1: SAAS vs. Conventional Machine Learning Models

Model category	Model	Micro-averaged precision	Micro-averaged recall	Micro-averaged F1-score	Micro-averaged HL
Paragraph vector-based (Doc2Vec)	LR	0.692***	0.524***	0.596***	0.091***
	SVM	0.768***	0.478***	0.589***	0.085***
	RF	0.700***	0.387***	0.499***	0.099***
	KNN	0.676***	0.454***	0.543***	0.097***
Term frequency-based (TF-IDF)	LR	0.763***	0.683***	0.721***	0.068***
	SVM	0.733***	0.670***	0.700***	0.073***
	RF	**0.846**	0.548***	0.546***	0.085***
	NB	0.778***	0.550***	0.645***	0.077***
	KNN	0.759***	0.626***	0.686***	0.073***
Proposed SAAS		0.807	**0.714**	**0.758**	**0.058**

Note: *, **, ***: Statistically significant difference at $p < 0.05, 0.01, 0.001$

Table 8. Example Segments SAAS Detected but Conventional ML Models Missed

Company	Segment	Row-wise attention weight	Ground truth	SAAS's predictions	TF-IDF + LR's predictions
Fool	when you place an order for a product or service we need to know the sort of information typically used for credit card transactions such as your name mailing and billing addresses and shipping address if different telephone number and credit card number and expiration date gathering this information allows us to process and fulfill your order and notify you of your order status we will also use your information to contact you regarding your order if necessary we encrypt all of this information using secure socket layers ssl technology	0.294	DS, FP	DS, FP	FP

The aspect with the highest row-wise attention weight (0.294) highlighted the phrase "using secure socket layers SSL technology," which likely belongs to "Data Security." While SAAS was able to capture security-related jargon using contextual information, term frequency-based and paragraph vector-based methods often suffered from the low frequency of each jargon phrase or word. A similar pattern occurred in 75 out of 1,872 segments in the test dataset (across all data practice categories). It is crucial to avoid missing identifying segments of data practices mandated to be disclosed by privacy regulations. For example, when a company encounters a data breach incident, regulators evaluate whether the company has sufficient security measures in place for it to avoid the incident and determine whether the business should take responsibility and be fined. In this example, if regulators were to use TF-IDF + LR, they might decide to levy unwarranted fines against the company for not incorporating content about appropriate security measures (i.e., secure sockets layer) when in fact, the company indeed included this information in the segment.

Experiment 2 Results: SAAS vs. Deep Learning Models

In Experiment 2, we compared the performance of SAAS against 10 state-of-the-art DL-based models. We present each model's performance in terms of micro-averaged precision, micro-averaged recall, micro-averaged F1-score, and micro-averaged HL in Table 9. The results are grouped based on their DL architecture. The best scores appear in boldface. Each model's performance in each of the 10 data practice categories is reported in Appendix B.

SAAS outperformed each benchmark method in terms of micro-averaged precision (0.807), micro-averaged F1-score (0.758), and micro-averaged HL (0.058) by statistically significant margins. SAAS consistently outperformed CNN (F1-score of 0.745) and unidirectional RNN-based methods (F1-scores between 0.730-0.749), indicating that operating in forward and backward directions captures more comprehensive local context information for distinguishing word semantics. In addition, SAAS attained a higher F1-score (0.758) than bidirectional RNN-based methods (F1-scores between 0.742 and 0.747), indicating that capturing global and local dependencies with the self-attention mechanism can disambiguate segment semantics (Du et al., 2020). SAAS outperformed the 10 independent SSASEs approach (F1-score of 0.749) and SSASE leveraging the multi-label classifier (F1-score of 0.750) by statistically significant margins. These results suggest that SAAS may have captured relationships and common features between data practice categories that each independent SSASE missed and differentiated aspects learned by the multi-head self-attention mechanism more effectively than the SSASE combined with a multi-label classifier.

Table 9. Experiment 2: SAAS vs. Deep Learning Models

Method category	Model	Micro-averaged precision	Micro-averaged recall	Micro-averaged F1-score	Micro-averaged HL
CNN-based	CNN	0.762***	0.729	0.745**	0.064***
Unidirectional RNN-based	LSTM + Max pooling	0.763***	0.719	0.739***	0.065***
	LSTM + Mean pooling	0.756***	0.706**	0.730***	0.067***
	GRU + Max pooling	0.773***	0.724	0.747**	0.063**
	GRU + Mean pooling	0.780**	0.721	0.749***	0.062***
Bidirectional RNN-based	BiLSTM + Max pooling	0.752***	**0.733**	0.742***	0.065***
	BiLSTM + Mean pooling	0.777***	0.712	0.743***	0.063***
	BiGRU + Max pooling	0.767**	0.728	0.746***	0.063***
	BiGRU + Mean pooling	0.779**	0.718	0.747**	0.063**
Attention-based	10 SSASEs, one for each data practice category	0.770**	0.729*	0.749***	0.062***
	SSASE with a multi-label classifier	0.800***	0.706*	0.750***	0.060***
	Proposed SAAS	**0.807**	0.714	**0.758**	**0.058**

Note: *, **, ***: Statistically significant difference at $p < 0.05, 0.01, 0.001$

Table 10. Example Segments SAAS Detected by SSASE Missed

Company	Segment	Row-wise attention weight	Ground truth	SAAS's predictions	SSASE's predictions
Fortune	these tracking technologies may be by us and or by our service providers or partners on our behalf these technologies enable us to assign a unique number to you and relate your service usage information to other information about you including your personal information we may match	0.2126	FP, TP	FP, TP	FP

We present an example segment that SAAS correctly classified as both FP and TP, but all benchmark approaches incorrectly classified as only FP or TP in Table 10. Instances related to FP and TP were selected because they had the highest number of labels (653 for FP and 548 for TP). The table presents the predicted data practice category generated by SAAS and SSASE with a multi-label classifier (best-performing benchmark method in terms of F1-score). The color shades represent the normalized word weights learned by the attention head that extracted the aspect of the segment with the highest row-wise attention weight. Darker shades indicate the higher importance of phrases in the segment. For illustration and clarity purposes, we set the number of attention heads as 5. If different aspects are treated equally, each aspect will have a row-wise attention weight of 0.2.

The aspect with the highest row-wise attention weight highlighted the phrase "by us and or by our service providers or partners." This phrase indicates that both the first party and third party would collect/access users' data. In addition, since FP and TP share common information, such as the types of collected personal information, segments that belong to FP (or TP) are more likely to also belong to TP (or FP). SAAS can leverage differentiating aspects and the relationship between data practice categories to identify both TP and FP, whereas SSASE only identified FP. A similar pattern occurred in 287 out of 1,872 segments in the testing dataset. Comprehensively capturing all labels is essential. In this example, if regulators were to evaluate annotated TP segments generated by SSASE, they would miss this segment and may impose unwarranted fines on the company for not documenting all relationships with third-party data processors. SAAS can help prevent unwarranted fines by returning the segment containing the information of third-party tracking technologies to confirm that the company did request consent in its privacy policy.

Experiment 3 Results: Ablation Analysis

In Experiment 3, we examined the effect of RWA and the multi-label classifier on SAAS's performance. Three variants of SAAS were tested, including SAAS without RWA, SAAS but replacing RWA with MLP, and 10 SAASs, one for each data practice category. We present each model's performance in terms of micro-averaged precision, recall, F1-score, and HL in Table 11. The best scores are highlighted in boldface. Each model's performance in each of the 10 data practice categories is reported in Appendix B.

Table 11. Experiment 3: Ablation Analysis

Model	Micro-averaged precision	Micro-averaged recall	Micro-averaged F1-score	Micro-averaged HL
Without RWA	0.800***	0.706***	0.750***	0.060***
Replacing RWA with MLP	0.802***	0.699***	0.747***	0.060***
10 SAASs, one for each data practice category	0.792*	0.659**	0.718***	0.066***
SAAS	**0.807**	**0.714**	**0.758**	**0.058**

Note: *, **, ***: Statistically significant difference at $p < 0.05, 0.01, 0.001$

SAAS outperformed the SAAS variants in terms of micro-averaged precision (0.807), micro-averaged recall (0.714), micro-averaged F1-score (0.758), and micro-averaged HL (0.058) by statistically significant margins. The results suggest that the RWA emphasized the critical aspects corresponding to each segment extracted by the multi-head self-attention mechanism to help improve performance. Replacing RWA with MLP resulted in a lower F1-score (0.747) than the proposed SAAS (0.758). An MLP layer can learn a fixed weighting matrix while RWA can update weights based on different input segments. The proposed SAAS, which simultaneously predicts 10 data practice categories, outperformed 10 independent SAAS models, possibly because the generated representation is aware of relationships and common features across categories.

Experiment 4 Results: Ground Truth vs. Automated Segmentation

In OPP-115, privacy policies were segmented manually. However, scaling this manual segmentation process to handle a large volume of privacy policies in practical applications is neither feasible nor efficient. In Experiment 4, we assessed SAAS's performance on two different segmentation approaches: ground truth segments from OPP-115 and automated segments. For the automated segmentation, we chose ST-Ro (Aumiller et al., 2021), as it outperformed prevailing text segmentation algorithms (e.g., GraphSeg, averaging over Global Vectors) on OPP-115 using the P_k metric (Beeferman et al., 1999) by statistically significant margins ($p < 0.001$). ST-Ro (Aumiller et al., 2021) operates by taking two neighbor sentences as input and predicting whether they belong to the same segments. In the 5×2 cross-validation, we trained SAAS on the first fold of the OPP-115 segmentation data in each round and used the second fold for testing. To generate ST-Ro segments, we concatenated neighboring sentences in the second fold if ST-Ro predicted that they should be part of the same segment. We also assigned data practice category labels to the ST-Ro segments if two or more of the original OPP-115 annotators indicated a specific data practice category within the segment.

The micro-averaged precision, recall, and F1-score for the original OPP-115 segmentation were 78.7%, 70.9%, and 74.6%, respectively. In comparison, the micro-averaged precision, recall, and F1-score for the ST-Ro segments were 76.3%, 71.2%, and 73.7%, respectively. The difference in F1-scores between these two segmentation methods was not statistically significant. These results indicate that SAAS is consistent in capturing key information for prediction when applied to both the original OPP-115 segmentation and the segments generated by the automated ST-Ro method. This suggests that while automated segmentation approaches may produce segments differently from manual methods, they do not significantly impact SAAS's overall performance.

Regulation Impact Detection: A Case on GDPR

To demonstrate proof of concept and the potential practical value of our proposed SAAS, we conducted a GDPR impact detection analysis on Amazon's privacy policies. While our proposed framework can be applied to any privacy regulation (new, updates to existing policies, or updated privacy policies), we focused the analysis on the impacts of GDPR on privacy policy evolution. GDPR was chosen because it impacts companies worldwide (since it protects all EU residents) rather than regionally (e.g., CCPA in California), and it has more documented global impacts than other recent regulations. We chose Amazon because it was recently fined $888 million based on accusations of using user data for developing targeted ads without attaining the consent of its users (thereby violating GDPR) (Dumiak, 2021). Therefore, our case study aims to identify whether Amazon's pre- and post-GDPR privacy policies provided information about the ad targeting system and if they explicitly asked users to agree to Amazon's use of their data (i.e., regulation impact detection). We employed five steps to execute the case study, which can be adopted by relevant stakeholders (e.g., legislators, regulators, researchers) in their privacy policy analysis.

Table 12. Selected Segments Pertaining to FP in Amazon's Privacy Policy Pre- and Post-GDPR

Time	Segment	Row-wise attention weight
Pre-GDPR (March 3, 2014)	what about cookies cookies are unique identifiers that we transfer to your device to enable our systems to recognize your device and to provide features such as click purchasing recommended for your personalized on other web sites amazon associate with content served by and web sites using checkout by amazon payment service and storage of items in your shopping cart between	0.2766
Post-GDPR (February 12, 2021)	what about cookies and other identifiers to enable our systems to recognize your browser or device and to provide and improve amazon services we use cookies and other identifiers for more information about cookies and how we use them please read our cookies notice	0.3044

Step 1: Collect privacy policies before and after a time of interest. GDPR became enforceable on May 25, 2018. Therefore, we collected Amazon's pre- and post-GDPR privacy policies from March 3, 2014, through February 12, 2021.

Step 2: Pre-process the privacy policies by dividing them into semantically coherent segments with text segmentation techniques. Consistent with the best practices (Fawaz et al., 2019; Harkous et al., 2018; Zaeem & Barber, 2021), we segmented the retrieved policies with automated text segmentation techniques, ST-Ro (Aumiller et al., 2021).

Step 3: Annotate segments with SAAS. We annotated segments in pre- and post-GDPR privacy policies using SAAS pre-trained on the OPP-115 corpus.

Step 4: Select data practice categories of interest. We selected segments labeled as FP for further investigation in this case study. FP contains user data information, including personally identifiable information and behavioral data that facilitate targeted ads.

Step 5: Visualize segments in data practice categories of interest using attention weights. To identify the difference in privacy policies and the presentation of data practice categories, we visualized segments using the feature weights assigned by SAAS's attention mechanisms.

We present two FP segments in Amazon's pre- and post-GDPR privacy policies in Table 12. The color shades are the normalized row-wise attention weights. Dark red indicates the higher importance of phrases in the segment. For illustration purposes, we set the number of attention heads as 5. If aspects are treated equally, each aspect will have a row-wise attention weight of 0.2.

As shown in Table 12, the segments are related to cookies, commonly known for collecting users' behavioral data to generate personalized recommendations. In the pre-GDPR privacy policy, the aspect with the highest row-wise attention weight (0.2766) highlighted examples of specific features that utilized cookies in such a segment. In the post-GDPR policy, the aspect with the highest row-wise attention weight (0.3044) highlighted the phrase that leads users to another document called "Cookies Notice." This redirect could increase the burden on users aiming to understand data practices related to targeted ads. Recitals 39 and 58 of the GDPR mandate transparency in data practices, i.e., require that any information addressed to the public or the data subject be concise, easily accessible, and easy to understand. However, in Amazon's updated policy, Amazon does not explicitly present the information about cookies. This information is hidden in a separate document that users have to locate, access, and read (thereby violating the "easily accessible" principle of GDPR). Users who cannot access or comprehend the separate document will not know what they are consenting to. Taken together, these results can help regulators identify how Amazon is adhering to GDPR in a targeted manner.

Discussion and Contributions

The increasing societal concern about consumer information privacy has led to new privacy regulations and fundamental changes in companies' privacy policies. Consequently, there is a need to evaluate how companies change their privacy policies and whether they provide more protection for users' information. In this study, we adopted the computational design science paradigm to design, implement, and evaluate a novel privacy analytics framework with a novel DL-based SAAS text analytics method that was guided by key privacy analytics domain requirements. SAAS automatically annotates segments in privacy policies into data practice categories. We rigorously evaluated SAAS against benchmark ML and DL methods and demonstrated its practical utility with an in-depth case study of

GDPR's impact on Amazon's privacy policies. Our research contributes to the IS knowledge base and has managerial and practical implications.

Contributions to the IS Knowledge Base

Novel IT artifacts often contribute prescriptive knowledge back to the IS knowledge base to guide future research (Hevner et al., 2004; Nunamaker et al., 1990; Rai, 2017; Zhu et al., 2021). Common contributions can include a situated implementation of an IT artifact in a selected domain and/or design principles that can be applied to other application environments. Our proposed privacy analytics framework is a situated implementation aligned with information privacy and data analytics. It also follows two key design principles that are applicable beyond the privacy policy evolution analysis: (1) differentiating the importance of different sets of words or phrases in a given complex and long text and incorporating common features between labels into a multi-label classification model for improving the model performance and (2) automatically annotating long text into finer-grained categories to facilitate downstream analytical tasks. Each design principle could help guide the design of IT artifacts for e-commerce, health, and privacy. Table 13 summarizes the framework components, the framework's general design principle, the relevant IS literature to which each principle could offer value, and potential classes of research inquiry. We then elaborate on how these design principles can provide value to each listed body of IS literature.

E-Commerce

Customer reviews in e-commerce help reveal relationships between users' preferences and product choices (Wu et al., 2019). Customer reviews often include complaints about, for example, the price, quality of the after-sales service, and other characteristics of a particular product. However, multi-class classification approaches cannot effectively identify and label product reviews that mention multiple preferences and products (Abrahams et al., 2015; Zhou et al., 2018). Scholars could consider Design Principle 1 to help design a multi-label classification approach to point out multiple interdependent (i.e., correlated) product issues (labels) with similar features mentioned in a customer review.

Health

Social media has become a popular channel for patients seeking health support (Bardhan et al., 2020). Users often post about multiple symptoms and afflictions (e.g., stress, physical disorders, mental disorders, etc.) in a single social media post (Chau et al., 2020). Scholars could consider Design Principle 1 when developing their multi-label classification approaches to automatically identify multiple (potentially related) symptoms or health needs in the content posted by patients to automatically group and label social media posts for further investigation.

Privacy

Regulations such as GDPR and CCPA have significant and lasting global impacts on developing legal documents. However, the length and complex nature of legal documents often make fundamental privacy analysis tasks (e.g., compliance checking) challenging to execute. Consequently, there is a significant need to develop approaches that can automatically synthesize the content in legal documents (e.g., Terms of Use) into manageable components. To this end, scholars could consider including Design Principle 2 when segmenting legal documents into coherent and semantically related sections that end users can select for their tasks.

Managerial and Practical Implications

Regulators and companies are increasingly focusing on the protection of consumer information privacy. Privacy policies are essential reference documents in examining how companies handle personal data. However, as businesses are affected by new privacy regulations, the complexity of privacy policies to be analyzed and reviewed has steadily increased (Amos et al., 2021). Our proposed framework automatically synthesizes the rich text content in privacy policies into semantically coherent data practice category label(s). We believe that our privacy analytics framework will help the two types of stakeholders in practice: regulators and companies.

Regulators: Our proposed framework can help identify content changes in different data practice categories. By analyzing how companies adjusted their privacy policies based on new or updated regulations, regulators could potentially improve their testing and evaluation processes for regulatory compliance and, therefore, enforce regulations more effectively.

Companies: Privacy policies will continue to change based on the functions provided by the business, requirements of new privacy regulations, and evolving consumer privacy expectations. Companies could use our proposed framework to review potential compliance issues in their privacy policies. Our framework can also help pinpoint specific data practice segments according to new (or updated) domestic and global privacy regulations. Consequently, companies could use our framework to better identify potential privacy violations and the associated legal and financial consequences.

Table 13. Design Principles Offered by our Proposed Privacy Policy Evolution Analytics Framework for Selected Bodies and Classes of IS Research Inquiry

Framework component	General design principle	Relevant IS literature	Potential class of research inquiry
SAAS	Differentiating the importance of different sets of words or phrases in a given complex and long text and incorporating common features and relationships between labels for improving multi-label classification	E-commerce	• Identifying and labeling multiple product issues within customer reviews
		Health	• Identifying patients' needs for social support from social media posts
Regulation impact	Automatically annotating long text into finer-grained categories to facilitate downstream analytical tasks.	Privacy	• Analysis of legal articles • Analysis of Terms of Use or End-User License Agreement documents

Conclusion and Future Directions

Consumer information privacy has rapidly emerged as a significant societal issue. Increasingly, legislators, regulators, and citizens are expressing concern about how companies maintain the information privacy of their consumers. Consequently, the past half-decade has seen the development and enforcement of various privacy regulations such as the EU's GDPR and the CCPA. However, many privacy policies that have been revised due to regulatory requirements have become increasingly lengthy and complex. Consequently, regulators and legislators often find it difficult to systematically identify how a company is amending and presenting its privacy policies.

In this study, we developed a novel privacy policy evolution analytics framework to help identify how companies change and present their privacy policies based on new privacy regulations. The core novelty of this framework is a SAAS method that automatically labels paragraph-length segments in long and unstructured privacy policies into their appropriate data practice category(ies) to help stakeholders focus on data practices of interest (without reading all of the text within privacy policies). SAAS incorporates RWA into the conventional SSASE to emphasize the differentiating features within segments during the data practice category labeling process. SAAS outperformed conventional ML approaches and state-of-the-art DL algorithms in conducting multi-label data practice segment annotation. We also illustrated SAAS's potential practical value with a case study identifying the differences between Amazon's privacy policies pre- and post-GDPR. The results of this case study indicate that Amazon's revised privacy policy requires consumers to exert more effort to find all the information related to targeted ads (violating a fundamental principle of GDPR). Regulators and legislators could leverage the proposed framework to amend their regulations to better protect consumers' information and to help companies evaluate the potential noncompliance of their privacy policies.

There are several promising directions for future research. First, different cultures and countries may address consumer privacy issues differently. To this end, future work could develop a multilingual privacy policy evolution analytics framework to handle multiple languages in different policy data categories and segments to investigate a privacy policy's global impact. Second, a SAAS-based AI assistance system with a user-friendly interface and browse, search, and recommendation functions could assist end users in large-scale online privacy policy comparisons from different companies when selecting online products or services. Each direction could help build a better understanding of how organizations and consumers respond to privacy policy requirements in a rapidly changing digital world.

Acknowledgments

We are grateful to the senior editor, associate editor, and the three anonymous reviewers for their constructive comments and feedback. This material is based upon work supported by the National Science Foundation under Grant OAC-2319325 and CNS-1936370.

References

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, *24*(6), 975-990. https://doi.org/10.1111/poms.12303

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, *30*(4), 736-758. https://doi.org/10.1002/jcpy.1191

Adjerid, I., Peer, E., & Acquisti, A. (2018). Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *MIS Quarterly*, *42*(2), 465-488. https://doi.org/10.25300/MISQ/2018/14316

Akanfe, O., Valecha, R., & Rao, H. R. (2020a). Assessing country-level privacy risk for digital payment systems. *Computers and*

Security, *99*. https://doi.org/10.1016/j.cose.2020.102065

Akanfe, O., Valecha, R., & Rao, H. R. (2020b). Design of an Inclusive Financial Privacy Index (INF-PIE): A financial privacy and digital financial inclusion perspective. *ACM Transactions on Management Information Systems*, *12*(1), 1-21. https://doi.org/10.1145/3403949

Alabduljabbar, A., Abusnaina, A., Meteriz-Yıldıran, Ü., & Mohaisen, D. (2021). TLDR: Deep learning-based automated privacy policy annotation with key policy highlights. In *Proceedings of the 20th Workshop on Privacy in the Electronic Society* (pp. 103-118). https://doi.org/10.1145/3463676.3485608

Alagha, I. (2022). Leveraging knowledge-based features with multilevel attention mechanisms for short Arabic text classification. *IEEE Access*, *10*, 51908-51921. https://doi.org/10.1109/ACCESS.2022.3175306

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., & Mayer, J. (2021). Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference* (pp. 2165-2176). https://doi.org/10.1145/3442381.3450048

Andow, B., Mahmud, S. Y., Wang, W., Whitaker, J., Enck, W., Reaves, B., Singh, K., & Xie, T. (2019). Policylint: Investigating internal privacy policy contradictions on Google Play. In *Proceedings of the 28th USENIX Security Symposium* (pp. 585-602).

Arora, S., Hosseini, H., Utz, C., Bannihatti, V. K., Dhellemmes, T., Ravichander, A., Story, P., Mangat, J., Chen, R., Degeling, M., Norton, T. Hupperich, T. Wilson, S., & Sadeh N. (2022). A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference* (pp. 5460-5472). https://aclanthology.org/2022.lrec-1.585

Aumiller, D., Almasian, S., Lackner, S., & Gertz, M. (2021). Structural text segmentation of legal documents. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law* (pp. 2-11). https://doi.org/10.1145/3462757.3466085

Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly*, *44*(1), 185-200. https://doi.org/10.25300/MISQ/2020/14644

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, *34*(1), 177-210. https://doi.org/10.1023/a:1007506220214

Bhatia, J., Breaux, T. D., Reidenberg, J. R., & Norton, T. B. (2016). A theory of vagueness and privacy risk perception. In *Proceedings of the IEEE 24th International Requirements Engineering Conference* (pp. 26-35). https://doi.org/10.1109/RE.2016.20

Bhatia, J., Evans, M. C., Wadkar, S., & Breaux, T. D. (2016). Automated extraction of regulated information types using hyponymy relations. In *Proceedings of the IEEE 24th International Requirements Engineering Conference Workshops* (pp. 19-25). https://doi.org/10.1109/REW.2016.22

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bowers, J., Reaves, B., Sherman, I., Traynor, P., & Butler, K. (2017). Regulators, mount up! Analysis of privacy policies for mobile money services. In *Proceedings of the 13th Symposium on Usable Privacy and Security* (pp. 97-114).

Breward, M., Hassanein, K., & Head, M. (2017). Understanding consumers' attitudes toward controversial information technologies: A contextualization approach. *Information Systems Research*, *28*(4), 760-774. https://doi.org/10.1287/isre.2017.0706

Buckman, J. R., Bockstedt, J. C., & Hashim, M. J. (2019). Relative privacy valuations under varying disclosure characteristics. *Information Systems Research*, *30*(2), 375-388. https://doi.org/10.1287/isre.2018.0818

Cao, Z., Hui, K. L., & Xu, H. (2018). An economic analysis of peer disclosure in online social communities. *Information Systems Research*, *29*(3), 546-566. https://doi.org/10.1287/isre.2017.0744

Cavusoglu, H. H., Phan, T. Q., Cavusoglu, H. H., & Airoldi, E. M. (2016). Assessing the impact of granular privacy controls on content sharing and disclosure on Facebook. *Information Systems Research*, *27*(4), 848-879. https://doi.org/10.1287/isre.2016.0672

Chang, C., Li, H., Zhang, Y., Du, S., Cao, H., & Zhu, H. (2019). Automated and personalized privacy policy extraction under GDPR consideration. In *Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications* (pp. 43-54). https://doi.org/10.1007/978-3-030-23597-0_4

Chau, M., Li, T. M., Wong, P. W., Xu, J. J., Yip, P. S., & Chen, H. (2020). Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly*, *44*(2), 933-955. https://doi.org/10.25300/MISQ/2020/14110

Chen, Y., Hu, D., Li, M., Duan, H., & Lu, X. (2022). Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. *International Journal of Medical Informatics*, *159*. https://doi.org/10.1016/j.ijmedinf.2021.104676

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103-111). https://doi.org/10.3115/v1/w14-4012

Cichy, P., Salge, T. O., & Kohli, R. (2021). Privacy concerns and data sharing in the internet of things: Mixed methods evidence from connected cars. *MIS Quarterly*, *45*(4), 1863-1892. https://doi.org/10.25300/MISQ/2021/14165

Crossler, R. E., & Bélanger, F. (2019). Why would I use location-protective settings on my smartphone? Motivating protective behaviors and the existence of the privacy knowledge-belief gap. *Information Systems Research*, *30*(3), 995-1006. https://doi.org/10.1287/isre.2019.0846

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2019). We value your privacy.. now take some cookies: Measuring the GDPR's impact on web privacy. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. https://doi.org/10.1007/s00287-019-01201-1

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(1), 1-30.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895-1923. https://doi.org/10.1162/089976698300017197

Du, M., Liu, N., & Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77. https://doi.org/10.1145/3359786

Dumiak, M. (2021). *Amazon faces $888M fine for GDPR violations*. JDSUPRA. https://www.jdsupra.com/legalnews/amazon-faces-888m-fine-for-gdpr-3663534/

Ebrahimi, M., Nunamaker, J. F., & Chen, H. (2020). Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach. *Journal of Management Information Systems, 37*(3), 694-722. https://doi.org/10.1080/07421222.2020.1790186

Evans, M. C., Bhatia, J., Wadkar, S., & Breaux, T. D. (2017). An evaluation of constituency-based hyponymy extraction from

privacy policies. In *Proceedings of the IEEE 25th International Requirements Engineering Conference* (pp. 312-321). https://doi.org/10.1109/RE.2017.87

Fawaz, K., Linden, T., & Harkous, H. (2019). Invited paper: The applications of machine learning in privacy notice and choice. In *Proceedings of the 11th International Conference on Communication Systems and Networks* (pp. 118-124). https://doi.org/10.1109/COMSNETS.2019.8711280

Fazzini, K. (2019). *Europe's sweeping privacy rule was supposed to change the internet, but so far it's mostly created frustration for users, companies, and regulators.* CNBC. https://www.cnbc.com/2019/05/04/gdpr-has-frustrated-users-and-regulators.html

Gal-Or, E., Gal-Or, R., & Penmetsa, N. (2018). The role of user privacy concerns in shaping competition among platforms. *Information Systems Research*, *29*(3), 698-722. https://doi.org/10.1287/isre.2017.0730

Galassi, A., Lippi, M., & Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 4291-4308. https://doi.org/10.1109/TNNLS.2020.3019893

Glavas, G., Nanni, F., & Ponzetto, S. P. (2016). Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics* (pp. 125-130). https://doi.org/10.18653/v1/s16-2016

Gopal, R. D., Hidaji, H., Patterson, R. A., Rolland, E., & Zhdanov, D. (2018). How much to share with third parties? User privacy concerns and website dilemmas. *MIS Quarterly*, *42*(1), 143-164. https://doi.org/10.25300/MISQ/2018/13839

Gopinath, A. A. M., Wilson, S., & Sadeh, N. (2018). Supervised and unsupervised methods for robust separation of section titles and prose text in web documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 850-855). https://doi.org/10.18653/v1/d18-1099

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proceedings of the 27th USENIX Security Symposium* (pp. 531-548).

Heimbach, I., & Hinz, O. (2018). The impact of sharing mechanism design on content sharing in online social networks. *Information Systems Research*, *29*(3), 592-611. https://doi.org/10.1287/isre.2017.0738

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105. https://doi.org/10.2307/25148625

Joty, S., Carenini, G., & Ng, R. T. (2013). Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, *47*, 521-573. https://doi.org/10.1613/jair.3940

Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering* (Article 14). https://doi.org/10.1145/3209280.3209526

Kerinec, E., Søgaard, A., & Braud, C. (2018). When does deep multi-task learning work for loosely related document classification tasks? In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. https://doi.org/10.18653/v1/w18-5401

Kim, S. H., & Kwon, J. (2019). How do EHRs and a meaningful use initiative affect breaches of patient information? *Information Systems Research*, *30*(4), 1184-1202. https://doi.org/10.1287/isre.2019.0858

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems*, *35*(2), 540-574. https://doi.org/10.1080/07421222.2018.1451957

Koh, B., Raghunathan, S., & Nault, B. R. (2017). Is voluntary profiling welfare enhancing? *MIS Quarterly*, *41*(1), 23-41. https://doi.org/10.25300/misq/2017/41.1.02

Krohn, J., Beyleveld, G., & Bassens, A. (2020). *Deep learning illustrated: A visual, interactive guide to artificial intelligence.* Addison-Wesley.

Kumar, V. B., Ravichander, A., Story, P., & Sadeh, N. (2019). Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *CEUR Workshop Proceedings: Privacy-Enhancing Artificial Intelligence and Language Technologies*.

Kumar, V., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., Schaub, F., & Sadeh, N. (2020). Finding a Choice in a Haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of the World Wide Web Conference* (pp. 1943-1954). https://doi.org/10.1145/3366423.3380262

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (pp. 2267-2273).

Leone, V., & Di Caro, L. (2020). The role of vocabulary mediation to discover and represent relevant information in privacy policies. In *Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems* (pp. 73-82). https://doi.org/10.3233/FAIA200851

Letarte, G., Paradis, F., Giguère, P., & Laviolette, F. (2018). Importance of self-attention for sentiment analysis. In *Proceedings of theEMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 267-275). https://doi.org/10.18653/v1/w18-5429

Li, X. B., & Qin, J. (2017). Anonymizing and sharing medical text records. *Information Systems Research*, *28*(2), 332-352. https://doi.org/10.1287/isre.2016.0676

Lin, Z., Feng, M., Dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*.

Linden, T., Khandelwal, R., Harkous, H., & Fawaz, K. (2020). The privacy policy landscape after the GDPR. In *Proceedings on Privacy Enhancing Technologies* (pp. 47-64). https://doi.org/10.2478/popets-2020-0004

Liu, F., Wilson, S., Schaub, F., & Sadeh, N. (2016). Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *Proceedings of the AAAI Fall Symposium (*Technical Report, FS-16-01-FS-16-05).

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-00a

Nejad, N. M., Graux, D., & Collarana, D. (2019). Towards measuring risk factors in privacy policies. In *Proceedings of the Workshop on Artificial Intelligence and the Administrative State* (pp. 18-20).

Nisal, N., Cherivirala, S. K., Sathyendra, K. M., Hagan, M., Schaub, F., Wilson, S., & others. (2017). Increasing the salience of data use opt-outs online. In *Proceedings of the 17th Symposium on Usable Privacy and Security*.

Nunamaker, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. *Journal of*

Management Information Systems, *7*(3), 89-106. https://doi.org/10.1080/07421222.1990.11517898

Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J., & Sadeh, N. (2018). PrivOnto: A semantic framework for the analysis of privacy policies. *Semantic Web*, *9*(2), 185-203. https://doi.org/10.3233/SW-170283

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É.(2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Poplavska, E., Norton, T. B., Wilson, S., & Sadeh, N. (2020). From prescription to description: Mapping the GDPR to a privacy policy corpus annotation scheme. *Frontiers in Artificial Intelligence and Applications*, *334*, 243-246. https://doi.org/10.3233/FAIA200874

Qamar, A., Javed, T., & Beg, M. O. (2021). Detecting compliance of privacy policies with data protection laws. *ArXiv Preprint ArXiv:2102.12362*. https://doi.org/10.48550/arXiv.2102.12362

Rai, A. (2017). Editor's comments: Diversity of design science research. *MIS Quarterly*, *41*(1), iii-xviii.

Ravichander, A., Black, A., Wilson, S., Norton, T., & Sadeh, N. (2019). Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 4947-4958). https://doi.org/10.18653/v1/d19-1500

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45-50). https://doi.org/10.13140/2.1.2393.1847

Sathyendra, K. M., Schaub, F., Wilson, S., & Sadeh, N. (2016). Automatic extraction of opt-out choices from privacy policies. In *Proceedings of the AAAI Fall Symposium* (Technical Report FS-16-04).

Sathyendra, K. M., Wilson, S., Schaub, F., Zimmeck, S., & Sadeh, N. (2017). Identifying the provision of choices in privacy policy text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 2774-2779). https://doi.org/10.18653/v1/d17-1294

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 145-158). https://doi.org/10.1007/978-3-642-23808-6_10

Slavin, R., Wang, X., Hosseini, M. B., Hester, J., Krishnan, R., Bhatia, J., Breaux, T. D., & Niu, J. (2016). Toward a framework for detecting privacy policy violations in Android application code. In *Proceedings of the International Conference on Software Engineering*. https://doi.org/10.1145/2884781.2884855

Spliethöver, M., Klaff, J., & Heuer, H. (2019). Is it worth the attention? A comparative evaluation of attention layers for argument unit segmentation. In *Proceedings of the 6th Workshop on Argument Mining* (pp. 74-82). https://doi.org/10.18653/v1/W19-4509

Stone, E. F., Gueutal, H. G., Gardner, D. G., & McClure, S. (1983). A field experiment comparing information-privacy values, beliefs, and attitudes across several types of organizations. *Journal of Applied Psychology*, *68*(3), 459-468. https://doi.org/10.1037/0021-9010.68.3.459

Story, P., Zimmeck, S., Ravichander, A., Smullen, D., Wang, Z., Reidenberg, J., Russell, N. C., & Sadeh, N. (2019). Natural language processing for mobile app privacy compliance. In *CEUR Workshop Proceedings: Privacy-Enhancing Artificial Intelligence and Language Technologies* (pp. 24-32).

Story, P., Zimmeck, S., & Sadeh, N. (2018). Which Apps Have Privacy Policies? In *Proceedings of the 6th Annual Privacy Forum* (pp. 3-23). https://doi.org/10.1007/978-3-030-02547-2_1

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018a). I read but don't agree: Privacy policy benchmarking using machine learning and the EU GDPR. In *Companion Proceedings of the Web Conference* (pp. 163-166). https://doi.org/10.1145/3184558.3186969

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018b). Privacyguide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the 4th ACM International Workshop on Security and Privacy Analytics, Co-Located with CODASPY 2018* (pp. 15-21). https://doi.org/10.1145/3180445.3180447

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1-13. https://doi.org/10.4018/jdwm.2007070101

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22-30. https://doi.org/10.1109/MCSE.2011.37

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., & Sadeh, N. (2016). The creation and analysis of a Website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1330-1340). https://doi.org/10.18653/v1/p16-1126

Wu, J., Huang, L., & Zhao, J. L. (2019). Operationalizing regulatory focus in the digital age: Evidence from an e-commerce context. *MIS Quarterly*, *43*(3), 745-764. https://doi.org/10.25300/MISQ/2019/14420

Wunderlich, P., Veit, D. J., & Sarker, S. (2019). Adoption of sustainable technologies: A mixed-methods study of German households. *MIS Quarterly*, *43*(2), 673-691. https://doi.org/10.25300/MISQ/2019/12112

Yin, W., Kann, K., Yu, M., & Schutze, H. (2017). *Comparative study of CNN and RNN for natural language processing*. arXiv. https://arxiv.org/pdf/1702.01923.pdf

Zaeem, R. N., & Barber, K. S. (2017). A study of web privacy policies across industries. *Journal of Information Privacy and Security*, *13*(4), 169-185. https://doi.org/10.1080/15536548.2017.1394064

Zaeem, R. N., & Barber, K. S. (2021). Comparing privacy policies of government agencies and companies: A study using machine-learning-based privacy policy analysis tools. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence* (Vol. 2, pp. 29-40). https://doi.org/10.5220/0010180700290040

Zalmanson, L., Oestreicher-Singer, G., & Ecker, Y. (2022). The role of social cues and trust in users' private information disclosure. *MIS Quarterly*, *46*(2), 1109-1134. https://doi.org/10.25300/MISQ/

2022/16288
Zhang, M. L., Li, Y. K., Liu, X. Y., & Geng, X. (2018). Binary relevance for multi-label learning: an overview. In *Frontiers of Computer Science*, *12*(2), 191-202. https://doi.org/10.1007/s11704-017-7031-7
Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., & Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems*, *35*(2), 510-539. https://doi.org/10.1080/07421222.2018.1451956
Zhu, H., Samtani, S., Brown, R., & Chen, H. (2021). A deep learning approach for recognizing activity of daily living (ADL) for senior care: Exploiting interaction dependency and temporal patterns. *MIS Quarterly*, *45*(2), 859-896. https://doi.org/10.25300/misq/2021/15574
Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Cameron Russell, N., & Sadeh, N. (2019). MAPS: Scaling privacy compliance analysis to a million apps. In. *Proceedings on Privacy Enhancing Technologies Symposium* (pp. 66-86). https://doi.org/10.2478/popets-2019-0037
Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N., Bellovin, S. M., & Reidenberg, J. (2017). Automated analysis of privacy requirements for mobile apps. In *Proceedings of the NDSS Symposium*. https://doi.org/10.14722/ndss.2017.23034

About the Authors

Fangyu Lin (fangyu.lin@utsa.edu) is an assistant professor in the Department of Information Systems and Cyber Security at the Carlos Alvarez College of Business, The University of Texas at San Antonio. She holds a Ph.D. in management information systems from the University of Arizona. Dr. Lin's research centers on artificial intelligence, deep learning, machine learning, information privacy analytics, and risk assessment. Her work has been published or accepted at a variety of prominent journals, conferences, and workshops, including *MIS Quarterly*, IEEE International Conference on Intelligence and Security Informatics, International Conference on Data Mining Workshops, International Conference on Computational Linguistics: System Demonstrations, MIPRO ICT and Electronics Convention, and Pacific Asia Conference on Information Systems. She has also contributed to multiple projects funded by the National Science Foundation.

Sagar Samtani (ssamtani@iu.edu) is an associate professor and Arthur M. Weimer Faculty Fellow in the Department of Operations and Decision Technologies and the founding executive director of the Data Science and Artificial Intelligence Lab at the Kelley School of Business at Indiana University (IU). He received his Ph.D. from the Artificial Intelligence (AI) Lab at the University of Arizona. Dr. Samtani's research focuses on developing AI-enabled algorithms and systems for cybersecurity, mental health, and business intelligence applications. He has published over 90 journal, conference, and workshop papers in *MIS Quarterly, Information Systems Research, Journal of Management Information Systems*, *IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Dependable and Secure Computing, ACM Transactions on Knowledge Discovery from Data, ACM Transactions on Privacy and Security*, and other outlets. His research has received over $7M funding from the NSF and other agencies. He has won several Best Paper awards for his research. Dr. Samtani has won the NSF CAREER Award, IU Outstanding Junior Faculty Award, the IEEE Big Data Security Junior Research Award, the AIS Early Career Award, the INFORMS Gordon B. Davis Young Scholar Award, the INFORMS Design Science Award, and the IU Trustees Teaching Award. He was inducted into the NSF/CISA CyberCorps SFS Hall of Fame and was named by Poets and Quants as a Top 50 Undergraduate Business School Professor. Dr. Samtani's work has received media attention from the AP, *WIRED, Forbes*, *Miami Herald*, Fox News, *Science Magazine*, and the AAAS.

Hongyi Zhu (hongyi.zhu@utsa.edu) is an assistant professor in the Department of Information Systems and Cyber Security at the Carlos Alvarez College of Business at The University of Texas at San Antonio. He received his Ph.D. in management information systems from the University of Arizona. Dr. Zhu's research focuses on developing advanced analytics for mobile and mental health, cybersecurity, and business intelligence. He has multidisciplinary research interests and has published in various journals, conferences, and workshops, including *MIS Quarterly, Journal of Management Information Systems, IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Privacy and Security, ACM Transactions on Management Information Systems, Journal of Biomedical Informatics*, and others. He is a member of the IEEE, ACM, AIS, and INFORMS.

Laura Brandimarte (lbrandimarte@arizona.edu) is an associate professor in the Department of Management Information Systems at the Eller College of Management at the University of Arizona. She received her Ph.D. in public policy and management from the Heinz College at Carnegie Mellon University. Her research focuses on the ethics of technology and behavioral aspects of privacy and security decision-making. Her work has been published in major academic journals, including *Science, Journal of Consumer Psychology, Journal of Experimental Psychology: General,* and *ACM Computing Surveys.*

Hsinchun Chen (hsinchun@arizona.edu) is Regents Professor and Thomas R. Brown Chair in Management and Technology in the Management Information Systems Department at the Eller College of Management, University of Arizona. He received his Ph.D. in information systems from New York University. He is the author/editor of over 20 books, 25 book chapters, 320 SCI journal articles, and 160 refereed conference articles covering web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management. He founded the AI Lab at The University of Arizona in 1989, which has received significant research funding ($60M+) from NSF, NIH, DOD, DOJ, CIA, DHS, and other agencies. He has served as editor-in-chief, senior editor or associate editor for major ACM/IEEE (*ACM Transactions on Information Systems, ACM Journal on Educational Resources in Computing, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Systems, Man, and Cybernetics*), MIS (*MIS Quarterly, Decision Support Systems*) and Springer (*Journal of the Association for Information Science and Technology*) journals and as conference/program chair for major ACM/IEEE/MIS conferences in digital library (ACM/IEEE JCDL, ICADL), information systems (ICIS), security informatics (IEEE ISI), and health informatics (ICSH). Dr. Chen is the director of the UA AZSecure Cybersecurity Program, with $10M+ funding from NSF SFS, SaTC, and CICI programs and CAE-CD/CAE-R cybersecurity designations from NSA/DHS. He is a fellow of ACM, IEEE, AIS, and AAAS.

Appendix A

SAAS Model Specifications

The proposed SAAS model is implemented with PyTorch (Paszke et al., 2019). Consistent with best practices in computational design science research, we provide the key architectural details and the parameter settings of our proposed model in Table A1 (Zhu et al., 2021).

Table A1. SAAS Model Specification

Component	Layer	Previous layer	Activation function	Dropout	Output shape
Input	Input	-	-	-	(500)
Word embedding	Embedding	Input	-	Yes	(500, 300)
Bi-LSTM	Bi-LSTM	Embedding	-	Yes	(500, 256*4)
Attention mechanism	Dense1	Bi-LSTM	tanh	No	(256*4, 512)
	Dense2	Dense1	Softmax	Yes	(512, 30)
Matrix sentence embedding	M_emb	Bi-LSTM, Dense2	-	-	(512, 30)
Row-wise attention	Dense3	M_emb	Tanh, Softmax	-	(30, 1)
Multi-label classifier	Dense4	M_emb	-	Yes	(512*30, 1024)
	Dense5	Dense4	Sigmoid	-	(1024, 1)
	Dense6	Dense4	Sigmoid	-	(1024, 1)
	⋮	⋮	⋮	-	⋮
	Dense14	Dense4	Sigmoid	-	(1024, 1)

To ensure the length consistency of the input data practice segments, we padded and truncated segments that contained fewer than or more than (respectively) 500 words. Each word was encoded as a 300-dimensional word embedding. Each Bi-LSTM direction contained two LSTM layers with a 256-dimensional hidden state. The four hidden states of Bi-LSTM were concatenated as the input of the attention mechanism and matrix sentence embedding. In the attention mechanism, the Dense1 layer generated 512 linear combinations of the Bi-LSTM hidden state, from which the Dense2 layer extracted 30 disparate aspects. Row-wise attention weighted the 30 aspects of matrix sentence embedding. Finally, matrix sentence embedding was passed to a dense layer of 1,024 nodes and 10 binary classifiers (i.e., Dense5 to Dense14) for multi-label classification. To evaluate the multi-label classification performance of the models, we used binary cross-entropy as the loss function. The Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0005 and a batch size of 32 was used to train SAAS. (Ebrahimi et al., 2020).

Appendix B

Performance Breakdown by Category

In the main text, we presented the results of the proposed SAAS and all benchmark machine learning (ML) and deep learning (DL) models across all 10 data practice categories. However, identifying how each approach is performed in each category can help stakeholders identify the appropriate model for particular categories. We present SAAS's performance against ML models by 10 data practice categories in Table B1. All models were evaluated based on precision, recall, F1-score, and hamming loss (HL). We performed paired *t*-tests to identify statistically significant differences between SAAS and the benchmark methods. The best scores appear in boldface.

Table B1. Performance of SAAS vs. Conventional ML Models by Category

Model category	Model	FP (*n*=1,522)	TP (*n*=1,186)	UCC (*n*=632)	UAED (*n*=231)	DR (*n*=156)	DS (*n*=375)	PC (*n*=192)	DNT (*n*=32)	ISA (*n*=353)	O (*n*=1,763)
Precision											
Paragraph vector-based (Doc2Vec)	LR	0.766**	0.718***	0.635**	0.550***	0.249***	0.641***	0.663***	0.400***	0.821***	0.616***
	SVM	0.783	0.757**	0.742	0.699**	0.100***	0.894	0.770***	0.583***	0.913	0.703***
	RF	0.726***	0.725***	0.610***	0.649**	0.081***	0.741*	0.871**	0.100***	0.810**	0.673***
	KNN	0.731***	0.713***	0.616***	0.615***	0.050***	0.684***	0.878**	0.667*	0.836***	0.591***
Term frequency-based (TF-IDF)	LR	0.787	0.781	0.679**	0.839	0.733	0.855	0.911	**1.000**	0.909	0.667***
	SVM	0.760**	0.742**	0.643***	0.828	0.723	0.848	0.902*	**1.000**	0.916	0.633***
	RF	**0.826**	**0.849**	**0.809**	**0.870**	0.818	**0.957**	**0.980**	0.700	**0.954**	**0.821**
	NB	0.767**	0.757**	0.760	0.780	0.794*	0.908	0.779***	0.982	0.909	0.759*
	KNN	0.742***	0.740***	0.700**	0.768	**0.823**	0.873	0.842**	**1.000**	0.898	0.748***
Proposed SAAS		0.799	0.802	0.787	0.791	0.798	0.851	0.933	0.974	0.907	0.802
Recall											
Paragraph vector-based (Doc2Vec)	LR	0.636***	0.556***	0.365***	0.422***	0.138	0.532***	0.669	0.547	0.728***	0.397***
	SVM	0.625***	0.523***	0.304***	0.291***	0.002**	0.429***	0.637**	0.528	0.670***	0.317***
	RF	0.538***	0.451***	0.221***	0.116***	0.006**	0.152***	0.442***	0.006***	0.287***	0.380***
	KNN	0.556***	0.473***	0.298***	0.126***	0.002**	0.262***	0.581***	0.077***	0.346***	0.551
Term frequency-based (TF-IDF)	LR	0.751***	0.711***	**0.537**	0.600	**0.323**	0.669	**0.785**	**0.903**	**0.817**	**0.625**
	SVM	0.738***	0.705***	**0.537**	0.580*	0.313	0.640**	0.735	0.858	0.771**	0.618
	RF	0.511***	0.458***	0.347***	0.085***	0.057*	0.350***	0.290***	0.110***	0.521***	0.369***
	NB	0.722***	0.558***	0.336***	0.260***	0.121	0.550***	0.632***	0.535	0.615***	0.475***
	KNN	0.750***	0.663***	0.445**	0.469***	0.168	0.554***	0.740	0.832	0.761***	0.534*
Proposed SAAS		**0.855**	**0.815**	0.499	**0.695**	0.121	**0.681**	0.732	0.468	0.813	0.585
F1-score											
Paragraph vector-based (Doc2Vec)	LR	0.694***	0.626***	0.461***	0.474***	0.175	0.579***	0.663***	0.453**	0.771***	0.481***
	SVM	0.695***	0.618***	0.428***	0.406***	0.004**	0.579***	0.694***	0.543	0.772***	0.436***
	RF	0.618***	0.556***	0.324***	0.193***	0.012**	0.250***	0.583***	0.012***	0.421***	0.485***
	KNN	0.631***	0.568***	0.400***	0.208***	0.004***	0.378***	0.697***	0.137***	0.488***	0.570***
Term frequency-based (TF-IDF)	LR	0.769***	0.744***	0.598	0.699	**0.441**	0.749	**0.842**	**0.948**	**0.859**	0.645*
	SVM	0.749***	0.722***	0.584*	0.681*	0.433	0.729**	0.809	0.923	0.836**	0.625***
	RF	0.631***	0.595***	0.343***	0.153***	0.106	0.511***	0.444***	0.172***	0.672***	0.509***
	NB	0.743***	0.642***	0.466***	0.388***	0.210	0.685**	0.697***	0.685	0.733***	0.584***
	KNN	0.746***	0.699***	0.543***	0.581***	0.274	0.677***	0.786*	0.907	0.824***	0.622***
Proposed SAAS		**0.825**	**0.806**	**0.603**	**0.737**	0.196	**0.752**	0.819	0.607	0.856	**0.675**
HL											
Paragraph vector-based (Doc2Vec)	LR	0.195***	0.180***	0.097***	0.040***	0.033***	0.045***	0.023***	0.011***	0.035***	0.249***
	SVM	0.191***	0.175***	0.093***	0.036***	0.025*	0.036***	0.019***	0.007**	0.032***	0.240***
	RF	0.232***	0.196***	0.105***	0.041***	0.026***	0.052***	0.021***	0.008***	0.064***	0.235***
	KNN	0.226***	0.195***	0.102***	0.041***	0.025*	0.050***	0.017***	0.008***	0.059***	0.243***
Term frequency-based (TF-IDF)	LR	0.158***	0.133***	0.082**	0.022	**0.020**	**0.026**	**0.010**	**0.001**	0.022	0.201***
	SVM	0.172***	0.147***	0.087***	0.023	**0.020**	0.028*	0.012	**0.001**	0.024**	0.216***
	RF	0.208***	0.169***	0.094***	0.040***	0.024	0.039***	0.024***	0.007***	0.041***	0.208***
	NB	0.173***	0.169***	0.088***	0.035	0.023	0.029*	0.018	0.004	0.036**	0.198***
	KNN	0.178***	0.155***	0.086***	0.029***	0.022	0.031***	0.013**	**0.001**	0.027***	0.189***
Proposed SAAS		**0.127**	**0.106**	**0.074**	**0.021**	0.023	**0.026**	0.011	0.005	**0.022**	**0.165**

SAAS outperformed other benchmarks on FP (0.825), TP (0.806), UCC (0.603), UAED (0.737), DS (0.752), and O (0.675) on F1-score (6 of the 10 categories) and on FP (0.127), TP (0.106), UCC (0.074), UAED (0.021), DS (0.026), ISA (0.022), and O (0.165) on HL. Furthermore, SAAS outperformed all conventional ML methods, except TF-IDF + RF, on FP (0.799), TP (0.802), UCC (0.787), DR (0.798), PC (0.933), and O (0.802) on precision. In addition, SAAS achieved the best recall on FP (0.855), TP (0.815), UAED (0.695), and DS (0.681). The results suggest that SAAS's use of the attention mechanism with Bi-GRU enabled the model to leverage the context information better to achieve higher classification performance than benchmark methods. However, SAAS did not outperform term frequency-based models on DR, PC, and DNT categories on all the metrics, likely due to a lack of training data. We also evaluated the performances of the DL-based benchmarks by data practice category. Table B2 summarizes model performances. The best scores are highlighted in boldface.

Table B2. Performance of SAAS vs. Prevailing Deep Learning Models by Data Practice Category

Model category	Model	FP (*n*=1,522)	TP (*n*=1,186)	UCC (*n*=632)	UAED (*n*=231)	DR (*n*=156)	DS (*n*=375)	PC (*n*=192)	DNT (*n*=32)	ISA (*n*=353)	O (*n*=1,763)
Precision											
CNN-based	CNN	0.809	0.783*	0.615***	0.674*	0.662	0.752**	0.795**	0.917	0.840***	0.755*
Uni-directional RNN-based	LSTM + Max pooling	0.823	0.823	0.614***	0.728	0.452	0.721***	0.842*	0.834*	0.860***	0.704**
	LSTM + Mean pooling	0.816	0.796	0.665**	0.663**	0.531	0.722***	0.851*	0.626*	0.878	0.722**
	GRU + Max pooling	0.826	**0.835**	0.664***	0.702**	0.540	0.727**	0.878*	0.890	0.855**	0.699***
	GRU + Mean pooling	0.801	0.817	0.675*	0.723*	0.571	0.751**	0.858*	0.924	0.884*	0.760
Bi-directional RNN-based	BiLSTM + Max pooling	**0.826**	0.798	0.645**	0.735*	0.460	0.789*	0.888*	0.965	0.865*	0.683***
	BiLSTM + Mean pooling	0.799	0.821	0.676**	0.793	0.692	0.743**	0.881*	0.839*	0.864**	0.734**
	BiGRU + Max pooling	0.820	0.815	0.636***	0.762	0.398*	0.741**	0.893*	0.883	0.893	0.695***
	BiGRU + Mean pooling	0.823	0.804	0.678**	0.750	0.551	0.761**	0.891*	0.960	0.878*	0.733**
Attention-based	10 SSASEs	0.797	0.795	0.662*	0.729*	0.565	0.745*	0.887	0.824	0.890	0.745
	SSASE with a multi-label classifier	0.787	0.811	0.758	**0.800**	0.748	**0.861**	0.924	0.930	0.876*	**0.825**
	Proposed SAAS	0.799	0.802	**0.787**	0.791	**0.798**	0.851	**0.933**	**0.974**	**0.907**	0.802
Recall											
CNN-based	CNN	0.776***	0.798	**0.648**	0.707	**0.355**	0.720	**0.829**	**0.832**	0.864	0.627
Uni-directional RNN-based	LSTM + Max pooling	0.798**	0.775*	0.617	0.649	0.181	**0.739**	0.703	0.569	0.832	0.638
	LSTM + Mean pooling	0.782**	0.786*	0.597	0.678	0.257	0.697	0.708	0.492	0.845	0.596
	GRU + Max pooling	0.808*	0.780	0.609	**0.716**	0.247	0.706	0.695	0.544	0.845	0.638
	GRU + Mean pooling	0.832	0.797	0.605	0.702	0.302	0.721	0.737	0.722	0.834	0.568
Bi-directional RNN-based	BiLSTM + Max pooling	0.788***	0.796	0.613	0.707	0.255	0.727	0.747	0.648	**0.873**	**0.662**
	BiLSTM + Mean pooling	0.811*	0.775**	0.591	0.669	0.355	0.733	0.703	0.615	0.823	0.587
	BiGRU + Max pooling	0.803**	0.791	0.632	0.667	0.166	0.715	0.726	0.805	0.841	0.643
	BiGRU + Mean pooling	0.805*	0.798	0.590	0.693	0.257	0.705	0.718	0.738	0.842	0.600
Attention-based	10 SSASEs	0.835	0.807	0.589	0.691	0.287	0.733	0.756	0.830	0.847	0.590
	SSASE with a multi-label classifier	0.835	0.789	0.550	0.684	0.155	0.676	0.727	0.736	0.838	0.555
	Proposed SAAS	**0.855**	**0.815**	0.499	0.695	0.121	0.681	0.732	0.468	0.813	0.585
F1-score											
CNN-based	CNN	0.791***	0.789**	0.628	0.683	0.445	0.731*	0.809	**0.859**	0.851	**0.684**
Uni-directional RNN-based	LSTM + Max pooling	0.809*	0.798	0.605	0.675	0.248	0.728**	0.750**	0.651	0.844	0.665
	LSTM + Mean pooling	0.797***	0.789*	0.624	0.665	0.318	0.701***	0.768*	0.462	0.860	0.649
	GRU + Max pooling	0.816*	0.805	0.630	0.705	0.331	0.711***	0.770*	0.643	0.848	0.664
	GRU + Mean pooling	0.814	0.805	**0.631**	0.705	0.387	0.729***	0.780*	0.787	0.856	0.648*
Bi-directional RNN-based	BiLSTM + Max pooling	0.805***	0.795*	0.621	0.718	0.268	**0.753**	0.806	0.732	0.867	0.670
	BiLSTM + Mean pooling	0.804***	0.797*	0.627	0.724	**0.450**	0.736**	0.775*	0.637	0.841*	0.648
	BiGRU + Max pooling	0.810**	0.801	0.626	0.710	0.221	0.721***	0.797	0.821	0.865	0.666
	BiGRU + Mean pooling	0.811*	0.799	0.623	0.716	0.331	0.725***	0.786	0.822	0.858	0.657
Attention-based	10 SSASEs	0.814*	0.798	0.614	0.702	0.366	0.731*	0.810	0.823	**0.867**	0.656
	SSASE with a multi-label classifier	0.806**	0.796	0.629	0.723	0.240	0.751	0.809	0.807	0.854	0.656
	Proposed SAAS	**0.825**	**0.806**	0.603	**0.737**	0.196	0.752	**0.819**	0.607	0.856	0.675
HL											
CNN-based	CNN	0.143***	0.116*	0.088**	0.028*	**0.021**	0.031**	0.013*	**0.002**	0.025**	0.169
Uni-directional RNN-based	LSTM + Max pooling	0.131	0.107	0.092**	0.027*	0.025	0.032***	0.015*	0.005	0.025*	0.187***
	LSTM + Mean pooling	0.138*	0.114	0.082*	0.030*	0.032	0.035***	0.014*	0.012	0.022	0.187***
	GRU + Max pooling	0.128	**0.102**	0.082*	0.026	0.023	0.034**	0.013*	0.004	0.025	0.188***
	GRU + Mean pooling	0.132	0.105	0.080*	0.025*	0.024	0.031***	0.014*	0.003	0.023	0.180***

Bi-directional RNN-based	BiLSTM + Max pooling	0.133	0.111	0.085**	0.024	0.043	0.028	0.012	0.003	0.022	0.190***
	BiLSTM + Mean pooling	0.137**	0.107	0.080*	0.022	0.022	0.031**	0.013*	0.005	0.025**	0.186***
	BiGRU + Max pooling	0.131	0.107	0.087*	0.024	0.025*	0.032**	0.012	0.003	**0.021**	0.188***
	BiGRU + Mean pooling	0.130	0.109	0.082*	0.024	0.024	0.031***	0.013*	0.003	0.023	0.183**
Attention-based	10 SSASEs	0.133	0.111	0.084*	0.025	0.023	0.032*	0.012	**0.002**	**0.021**	0.180***
	SSASE with a multi-label classifier	0.140	0.109	**0.073**	0.022	0.023	**0.026**	**0.011**	0.003	0.023	0.169
	Proposed SAAS	**0.127**	0.102	0.074	**0.021**	0.023	**0.026**	**0.011**	0.005	0.022	**0.165**

SAAS achieved the highest F1-score on FP (0.825), TP (0.806), UAED (0.737), and PC (0.819) and the lowest HL on FP (0.127), UAED (0.021), DS (0.026), PC (0.011), and O (0.165). SAAS achieved the best precision on UCC (0.787), DR (0.798), PC (0.933), DNT (0.974), and ISA (0.907). In addition, SAAS outperformed prevailing DL models on recall on FP (0.855) and TP (0.815). No other model attained the best performances in two or more categories in any metric. The results indicate that SAAS consistently considers the unique differentiating aspects of all data practice categories when annotating segments. We also examined how the proposed SAAS and its variant performed in each category, summarized in Table B3. The best scores appear in boldface.

Table B3. Performance of Ablation Analysis by Category

Model	FP (*n*=1,522)	TP (*n*=1,186)	UCC (*n*=632)	UAED (*n*=231)	DR (*n*=156)	DS (*n*=375)	PC (*n*=192)	DNT (*n*=32)	ISA (*n*=353)	O (*n*=1,763)
Precision										
Without RWA	0.787	0.811	0.758	0.800	0.748	**0.861**	0.924	0.930	0.876*	0.825
Replacing RWA with MLP	0.785	**0.818**	0.751	0.780	0.745	0.821	0.928	0.970	**0.916**	0.816
10 SAASs	0.780	0.812	0.712*	**0.802**	0.325*	0.761	**0.944**	0.919**	0.845**	**0.833**
SAAS	**0.799**	0.802	**0.787**	0.791	**0.798**	0.851	0.933	**0.974**	0.907	0.802
Recall										
Without RWA	0.835	0.789	**0.550**	0.684	**0.155**	0.676	0.727	**0.736**	**0.838**	0.555
Replacing RWA with MLP	0.840	0.779	0.542	0.610	0.111	**0.706**	**0.755**	0.671	0.786	0.551*
10 SAASs	0.786**	0.761**	0.528	0.625*	0.087	**0.706**	0.652	0.843	0.851	0.451***
SAAS	**0.855**	**0.815**	0.499	**0.695**	0.121	0.681	0.732	0.468	0.813	**0.585**
F1-score										
Without RWA	0.806**	0.796	**0.629**	0.723	**0.240**	0.751	0.809	0.807	0.854	0.656
Replacing RWA with MLP	0.810	0.796	0.625	0.655	0.185	**0.753**	**0.830**	0.783	0.842	0.655
10 SAASs	0.782***	0.782*	0.588	0.691*	0.133	0.719**	0.749	**0.855**	0.845	0.579***
SAAS	**0.825**	**0.806**	0.603	**0.737**	0.196	0.752	0.819	0.607	**0.856**	**0.675**
HL										
Without RWA	0.140*	0.109	**0.073**	0.022	**0.023**	**0.026**	0.011	0.003	0.023	0.169
Replacing RWA with MLP	0.138	0.108	0.074	0.028	**0.023**	0.027	**0.010**	0.003	0.024	0.169
10 SAASs	0.152***	0.114	0.083*	0.024	**0.023**	0.032*	0.013	**0.002**	0.026	0.189***
SAAS	**0.127**	**0.106**	0.074	**0.021**	**0.023**	**0.026**	0.011	0.005	**0.022**	**0.165**

SAAS outperformed its variants on the majority of data practice categories on F1-score (FP: 0.825; TP: 0.806; UAED: 0.737; ISA: 0.856; O: 0.675). This is mainly because the row-wise attention can emphasize the critical semantics in segment embedding extracted by the multi-head self-attention mechanism. In addition, the results indicated that the proposed row-wise attention operation contributes to performance improvement by leveraging the dynamic weighting. Furthermore, compared to the variant that leveraged 10 independent binary classification models, SAAS can capture the relationships and common features between data practice categories.

Appendix C

Sensitivity Analysis of Our Proposed SAAS

We examined SAAS's sensitivity to four key sets of DL parameters: the number of hidden states in Bi-LSTM, the number of attention units, the number of attention heads, and the number of units in the dense layer of the multi-label classifier. We compared model performances based on micro-averaged precision, micro-averaged recall, micro-averaged F1-score, and micro-averaged HL. The baseline SAAS had 256 Bi-LSTM hidden states, 256 attention units to extract aspects of segments into a 30-head matrix segment embedding, and 1024 units in the dense layer. We changed the target parameter and fixed all other parameters for each SAAS variation. All models were compared with the baseline model for each set of parameters to examine the statistical significance. In this setup, the null hypothesis assumes that there is no significant difference between each model and the baseline model. We summarize the performance of SAAS and its variants in Table C1. The best performance of each parameter for each metric appears in boldface.

Table C1. Performance of SAAS with Parameter Variations

Number of hidden states in Bi-LSTM (baseline model: 256 hidden states)				
Model	**Micro-averaged precision**	**Micro-averaged recall**	**Micro-averaged F1-score**	**Micro-averaged HL**
128 hidden states	0.806	0.714	0.757	**0.058**
256 hidden states	**0.807**	0.714	**0.758**	**0.058**
512 hidden states	0.796*	**0.722**	0.757	0.059
Number of attention units (baseline model: 256 attention units)				
128 units	0.806	0.713	0.756	0.059
256 units	0.807	**0.714**	**0.758**	**0.058**
512 units	**0.809**	**0.714**	**0.758**	**0.058**
Number of attention heads (baseline model: 30 heads)				
20 heads	0.801	**0.715**	0.755	0.059
30 heads	**0.807**	0.714	**0.758**	**0.058**
40 heads	0.800	0.711	0.752	0.060
Number of units in the dense layer of the multi-label classifier (baseline model: 1024 units)				
512 units	**0.807**	0.712	0.756	0.059
1024 units	**0.807**	**0.714**	**0.758**	**0.058**
2048 units	0.804	0.711	0.755	0.059

Note: *Statistically significant difference at $p < 0.05$

When the number of Bi-LSTM hidden states increased from 128 to 256, there was no significant difference in micro-averaged precision (between 0.806 and 0.807), micro-averaged recall (between 0.714 and 0.714), micro-averaged F1-score (between 0.757 and 0.758), or micro-averaged HL (between 0.058 and 0.058). Further, increasing the number of hidden states did not yield statistically significant performance differences in micro-averaged F1-score or micro-averaged HL. Similarly, altering the number of attention units, attention heads, and units in the dense layer did not affect the statistically significant differences for any performance metric. This suggests that SAAS performance was not sensitive to parameter changes on the attention unit, attention head, or unit-in-the-dense layer. In particular, the results of the changes in the number of attention heads suggest that only a few differentiating aspects of a segment are needed.