Automated Analysis of Changes in Privacy Policies: A Structured Self-Attentive Sentence Embedding Approach

ABSTRACT

The increasing societal concern for consumer information privacy has led to the enforcement of new global privacy regulations such as the EU General Data Protection Regulation (GDPR) in 2018 and the California Consumer Privacy Act (CCPA) in 2020. Increasingly, there is a need to evaluate how companies revise their privacy policies corresponding to regulations and whether they improve the protection of users' information. However, prevailing privacy policy analysis and IS information privacy studies mainly employ manual analysis or behavioral/economic methods that lack scalability. This study adopts the computational design science paradigm to extend a deep learning-based Structured Self-Attentive Sentence Embedding (SSASE) approach to measure privacy policy evolution quantitatively. The proposed Self-Attentive Annotation System (SAAS) can leverage differentiating attention heads to improve the performance of multi-label data practice annotation tasks, which can facilitate data practice category-level privacy policy evolution analysis. We rigorously evaluate the proposed approach against state-of-the-art machine learning and deep learning benchmark methods on a well-established privacy policy dataset. We also demonstrate its practical utility with an in-depth case study of GDPR's impact on Amazon's privacy policies. Beyond demonstrating the managerial and practical implications, we discuss several design principles that can guide future design science-based e-commerce, health, and privacy applications.

Keywords: privacy policy, structured self-attentive sentence embedding, deep learning, GDPR, privacy analytics, computational design science

Automated Analysis of Changes in Privacy Policies: A Structured Self-Attentive Sentence Embedding Approach

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 major societal issue (Kokolakis 2017). Increasing concern over how companies maintain the information privacy of their consumers has led to the development and enforcement of global privacy regulations such as the EU General Data Protection Regulation (GDPR) in 2018 and the California Consumer Privacy Act (CCPA) in 2020. Each privacy regulation stipulates how companies control their customer's personal information. Companies that violate regulations can incur significant financial fines and lose their reputation. Table 1 summarizes recent major instances of companies violating GDPR. Events are summarized based on the occurrence date, company name, industry type, country the company is located in, the fine incurred, and the nature of the violation.

Table 1. Selected Recent Cases of GDPR Violations

Date	Company	Industry Type	Country	Fine	Violation
2020.3	Google	Search Engine	Sweden	$8M	Right to be forgotten
2020.3	The Royal Dutch Lawn Tennis Association	Sports	Netherlands	$560K	Third Party Sharing/Consent
2020.1	DSG	Retail	UK	$580K	Safeguards
2019.11	Deutsche Wohnen	Property	Germany	$15K	Right to be forgotten
2019.10	Facebook	Social Media	UK	$582k	Third Party Sharing/Consent
2019.8	Morele.net	Retail	Poland	$689K	Safeguards
2019.7	Life at Parliament View Ltd	Estate	UK	$92K	Safeguards
2019.7	British Airways	Airline	UK	$230M	Safeguards
2019.7	Unicredit Bank	Banking	Romania	$137K	Safeguards
2019.6	EE	Telco	UK	$116K	Consent
2019.1	Google	Search Engine	France	$57M	Transparency

The scope, scale, and far-reaching implications of emerging global privacy regulations have motivated many companies to evaluate how details about their data practices (i.e., collecting,

processing, storing, sharing, and protecting customer data) are presented in their privacy policies. Increasingly, legislators, regulators, and researchers are carefully reviewing privacy policies pre- and post-GDPR to ascertain how a company complies with the new regulations (i.e., privacy policy evolution). However, many privacy policies have grown in length and become more difficult to comprehend after a regulation (Amos et al. 2021; Degeling et al. 2019; Linden et al. 2020). In Figure 1, we present Google's privacy policy pre-GDPR (February 25, 2015) and post-GDPR (January 22, 2019) to illustrate how the privacy policy grew in length and complexity.

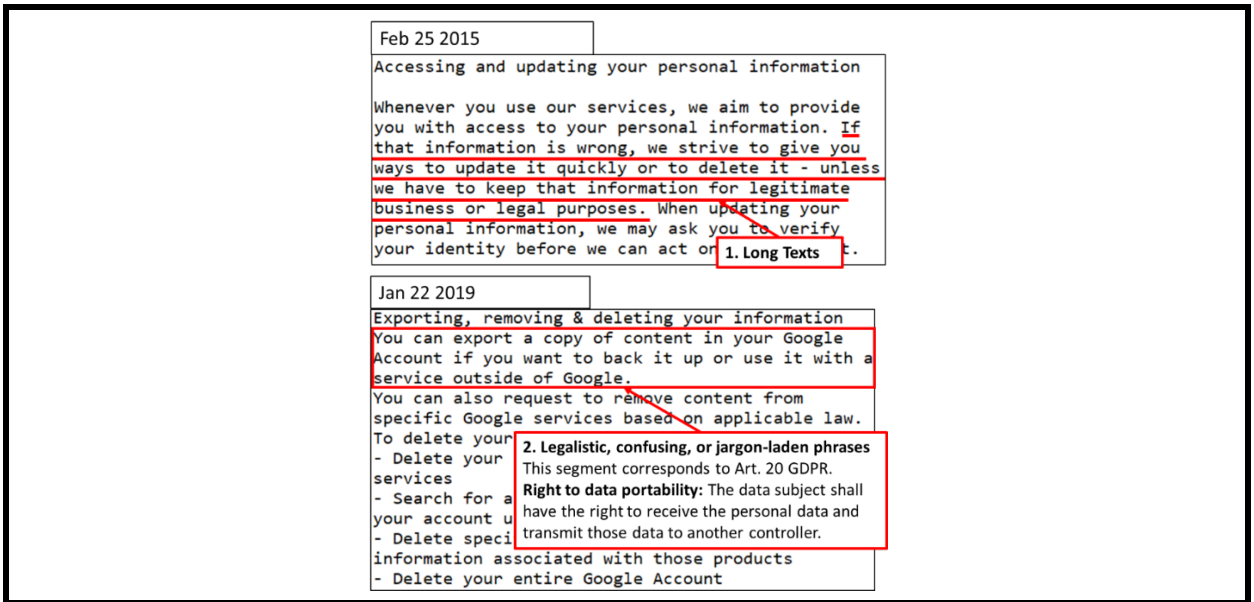


Figure 1. Google's Privacy Policy Before (Top) and After (Bottom) GDPR

The "Accessing and updating your personal information" section in Google's pre-PGDR privacy policy contained sentences that exceeded 30 words. Following the implementation of GDPR, the "Exporting, removing & deleting your information" section in the post-GDPR privacy policy contained legalistic or jargon-laden phrases that require college reading level to comprehend (Gluck et al. 2019). These characteristics can make tasks such as compliance checking non-trivial and challenging for regulators, companies, and other interested stakeholders

(Reidenberg et al. 2015). Consequently, there is a significant need to identify how the characteristics of privacy policies change (i.e., evolve) after a new regulation is implemented.

Despite widespread interest in identifying how privacy policies evolve due to regulatory demands, prevailing approaches use manual analysis to identify key differences (Degeling et al. 2019; Zaeem and Barber 2017). Such approaches are labor-intensive, time-consuming, and may be error-prone due to lengthy texts with confusing phrases (Gluck et al. 2019; Jain et al. 2016). While Information Systems (IS) scholars are uniquely positioned to analyze the text characteristics of privacy policies to identify how they evolve, prevailing IS privacy literature has focused on privacy concerns, risks, and controls from behavioral and economic perspectives. Methods within these paradigms were not designed to analyze the rich text within privacy policies. Consequently, 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 after a regulation is implemented.

In this study, we adopted the computational design science paradigm to design, develop, and evaluate a privacy policy evolution analytics framework that automatically analyzes long and complex privacy policy texts to help identify what text content changed after a regulation has been implemented. At the heart of this framework stands a novel DL-based Self-Attentive Annotation System (SAAS) that draws upon an emerging Structured Self-Attentive Sentence Embedding (SSASE) and attention mechanisms. SAAS automatically annotates segments in privacy policies with their data practice categories corresponding to prevailing privacy regulations (e.g., GDPR) and visualizes key changes before and after a regulation is implemented. 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 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, social media analytics, 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 identify research gaps within extant literature and pose research questions for study. Third, we detail the proposed privacy policy evolution analytics framework and its constituent components. Fourth, we present the results of our experiments and case study. Fifth, we discuss this study's contributions to the IS knowledge base and selected managerial implications. Finally, we summarize promising directions for future research and conclude this study.

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 the changes in companies' privacy policies. Third, we review SSASE to understand how a prevailing DL-based text analytics technique could be leveraged for automated privacy policy evolution analytics. Finally, attention mechanisms are examined to identify approaches that can dynamically weigh input features from privacy policies to enhance SSASE performance.

IS Information Privacy Research and the Computational Design Science Guidelines

Information privacy is fundamentally defined as "the ability of the individual to control personally 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 carefully examine varying aspects of information privacy. To date, IS scholars have leveraged behavioral, economic, and design science paradigms to make remarkable progress in four major categories of information privacy research: concern, control, risk, and preservation. In Table 2, we summarize the focus, IS paradigm, and the analytical method(s) leveraged in recent selected IS information privacy studies.

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

Category	Year	Author(s)	Focus	IS Paradigm	Analytical Method(s)
Privacy Concern	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 and Bélanger	Factors affecting the use of privacy settings on smartphones	Behavioral	Structural Equation Model
	2018	Adjerid et al.	Examining rational cognition and heuristics of privacy decision making	Behavioral	Linear Regression
	2018	Gopal et al.	Users' privacy concerns on 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, Structural Equation Model
	2017	Koh et al.	The impact of privacy cost on voluntary profiling in e-commerce platforms	Economic	Econometric model
Privacy Control	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 and Hinz	The impact of content provider platforms' sharing mechanism leveraged different privacy protection measures 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 and Kwon	The impact of EHRs and meaningful use on the risk of patient information breaches	Economic	Cox Proportional Hazards Model
Privacy Preservation	2017	Li and Qin	A novel privacy-preserving approach for unstructured health information	Design Science	Text Mining
	2016	Menon and Sarkar	Sanitizing sensitive information in large transactional databases	Design Science	Linear Programming

Research examining privacy concerns, controls, and risks primarily employed behavioral theories or econometric models to investigate the impact of privacy concerns on user or

organization decision-making, explore how various privacy controls influence users' behaviors, or examine the factors that affect privacy risk (Cao et al. 2018; Kim and Kwon 2019; Wunderlich et al. 2019). Despite their important contributions, the analytical methods (e.g., regression models) 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 how to design, develop, and evaluate 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.

IT artifacts developed under 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, Li and Qin (2017) in which unique data characteristics guided the development of a novel text analytics framework that incorporated carefully constructed feature representations and algorithms to anonymize medical records (Li and Qin 2017). 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 back to the IS knowledge base to help guide future related 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. In this study, this requires reviewing key data

characteristics of privacy policies and examining how advanced text analytics techniques can be developed to analyze them. Therefore, we review extant privacy policy analysis literature next.

Privacy Policy Analysis

A privacy policy is a legal contract binding a company in collecting, processing, sharing, and storing users' personal information (Antón et al. 2007). Privacy regulations (e.g., GDPR, CCPA) require companies to disclose their data practices in their privacy policies in a transparent manner to help users (e.g., customers) understand their rights to control their data. In general, ten major categories of data practices exist (Wilson et al. 2016). In Table 3, we describe each data practice category and specify selected recent regulations that require companies to disclose the category in their privacy policies.

Table 3. Summary of Key Data Practice Categories

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

First Party Collection (FP), Third Party Sharing/Collection (TP), Data Retention (DR), and Data Security (DS) detail what, how, and why first parties and third parties collect, process, store, share, and protect customer data. User Choice/Control (UCC), User Access, Edit, and Deletion (UAED), and Do Not Track (DNT) pertain to a user's rights. International & Specific Audiences 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 data practice categories (Wilson et al. 2016). Although recent privacy regulations clearly specify the requirements for information disclosure, there is no standard or recommended format for companies to 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.

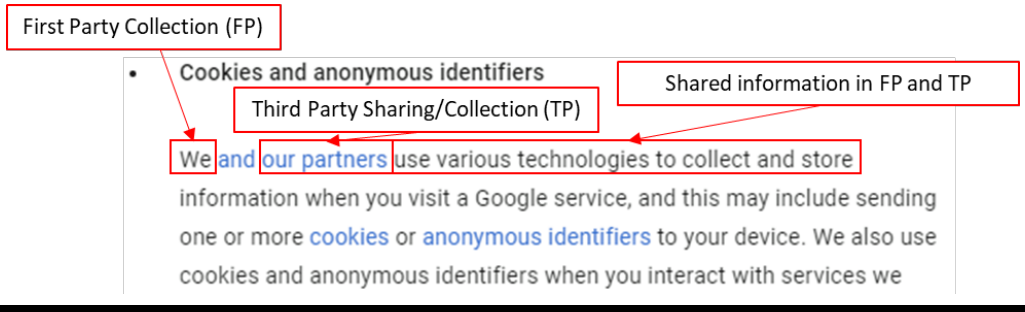


Figure 2. A Sample Segment in Google's Privacy Policy

In the "Cookies and anonymous identifiers" section of Google's privacy policy, the first segment contains details about FP (indicated by the word "We"), TP (indicated by the words "our partners"), and content pertaining to both FP and TP (indicated by the phrase "use various technologies to collect and store"). Dispersing and mixing data practice information in segments increases the difficulty of analyzing privacy policies (Degeling et al. 2019; Linden et al. 2020). However, privacy policies are currently the primary instruments stakeholders (e.g., consumers, regulators, companies) rely on to understand a companies' data practices (Amos et al. 2021). Therefore, we review selected recent privacy policy analysis research in Table 4. The summary is organized based on the focus of the study, the dataset examined, the corresponding privacy regulations, and the methodology employed. We also indicate whether a visualization system or interface is provided for end-users to examine the results of the analysis.

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**				**Visualization Component**
								Readability Metrics	**Keyword Searching**	**Descriptive Statistics**	**Machine Learning*****	
2021	Amos et al.	Comparative Study	English websites	1M	22	GDPR	No	Yes	Yes	Yes	No	Bar Chart, Line Chart
2021	Qamar et al.	Compliance Detection	OPP-115	115	1	GDPR, PDPA	No	No	No	No	BOW + LR, SVM, BERT	Bar Chart, Word Cloud
2021	Zaeem & Barber	Comparative Study	OPP-115	115	1	No	No	No	No	No	CNN, NB	Bar Chart
2020	Akanfe et al.	Privacy Risk Assessment	Mobile wallets and remittance apps	353	1	GDPR	No	No	Yes	No	No	Scatter plot
2020	Akanfe et al.	Privacy Risk Assessment	Mobile wallets and remittance apps	230	1	GDPR	No	No	Yes	No	BOW, LDA	Scatter plot
2020	Linden et al.	Comparative Study	OPP-115	115	2	GDPR	Yes	Yes	Yes	Yes	CNN	Bar Chart, Line Chart
2019	Andow et al.	Compliance Detection	Apps from Google Play Store	11K	1	No	No	No	Yes	No	Parse Tree + Rule-based	Bar Chart, Line Chart, Heatmap
2019	Chang et al.	Privacy Settings Assistant	OPP-115	115	1	GDPR	No	No	No	No	CNN, RF	Bar Chart
2019	Degeling et al.	Comparative Study	EU websites	112K	12	GDPR	Yes	No	Yes	Yes	No	Bar Chart, Line Chart
2019	Fawaz et al.	Comparative Study, Risk Assessment, Privacy Settings Assistant	OPP-115	115	2	GDPR	No	No	No	No	CNN	Bar Chart, Line Chart,
2019	Nejad et al.	Privacy Risk Assessment	OPP-115	115	1	GDPR	No	No	Yes	No	Did not specified	No
2019	Zimmeck et al.	Compliance Detection	Apps from Google Play Store	1M	1	GDPR, COPPA, CalOPPA	No	No	No	Yes	BOW + SVM	Bar Chart, Line Chart, Pie Chart, Heatmap
2018	Harkous et al.	QA System	OPP-115	115	1	No	No	No	No	No	CNN	Bar Chart

2018	Tesfay et al.	Privacy Risk Assessment	EU websites	45	1	GDPR	No	No	No	No	BOW + NB, SVM, DT, RF	No
2018	Story et al.	Comparative Study	Apps from Google Play Store	3M	3	CalOPPA, DOPPA, FIPPs	No	No	No	Yes	No	Bar Chart
2017	Zaeem and Barber	Comparative Study	NYSE, Nasdaq, and AMEX	600	1	FIPPs, COPPA	Yes	No	No	No	No	Bar Chart
2018	Oltramari et al.	QA System	OPP-115	115	1	No	No	No	No	No	Rule-based	Segment Coloring
2017	Sathyendra et al.	Privacy Settings Assistant	OPP-115	115	1	FIPPs	No	No	Yes	No	BOW, LDA, Parse Tree + LR	No
2016	Bhatia et al.	Compliance Detection	US websites	15	1	No	No	No	No	No	Parse Tree	No
2016	Sathyendra et al.	Privacy Settings Assistant	OPP-115	115	1	FIPPs	No	No	Yes	No	BOW + LR, SVM, RF, NB, KNN	No
2007	Antón et al.	Comparative Study	Health Institutions	24	1	HIPAA	Yes	Yes	No	No	No	No
2004	Anton et al.	Comparative Study	Financial Institutions	40	1	GLBA	Yes	Yes	No	No	No	No

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; GLBA: The Gramm-Leach-Bliley Act; 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; HMM: Hidden Markov Model; KNN: k-Nearest Neighbors; LDA: Latent Dirichlet Allocation; LR: Logistic Regression; NB: Naïve Bayes; RF: Random Forest; RNN: Recurrent Neural Network; SVM: Support Vector Machine; TF-IDF: Term Frequency-Inverse Document Frequency.

Major themes of extant privacy policy analysis literature include compliance detection (Andow et al. 2019; Qamar et al. 2021), privacy risk assessment (Akanfe et al. 2020a; Fawaz et al. 2019), privacy setting assistants (Chang et al. 2019; Sathyendra et al. 2016), and comparative analysis (Amos et al. 2021; Linden et al. 2020; Qamar et al. 2021). The prevailing dataset for conducting comparative analysis of privacy policies (the main focus of our study) is "Online Privacy Policies, set of 115 (OPP-115)" (Harkous et al. 2018). OPP-115 is a labeled privacy policy dataset developed by the Usable Privacy Policy Project (UPPP) from Carnegie Mellon University (CMU) that consists of 115 English privacy policies published between 2003 and 2015 from well-known, highly ranked websites based on Google trends across 15 sectors (as defined by DMOZ.org) (Wilson et al. 2016). Each privacy policy was manually partitioned into paragraph-length segments and assigned one or more data practice category labels by three law school students. Although the annotation scheme of OPP-115 is agnostic to particular laws, it reflects essential data practice categories provisioned by privacy regulations (e.g., GDPR) (Leone and Di Caro 2020; Poplavska et al. 2020).

With regards to methodology, many prior studies relied on manual analysis, keyword searching, or descriptive statistics (Degeling et al. 2019; Story et al. 2018; Zaeem and Barber 2017). Since privacy policies are long documents with no standard format, manual analysis can result in incomplete content extraction and have limited scalability. Studies employing automated techniques have trained ML algorithms (e.g., SVM, RF) on OPP-115 to classify segments in privacy policies into one or more data practice categories (i.e., multi-label classification) (Harkous et al. 2018). The goal of conducting classification in this fashion is to annotate each segment with one or more labels (i.e., data practice category) to help facilitate subsequent targeted analyses about specific components of a privacy policy (e.g., identify how a specific data practice category

changed). However, conventional ML algorithms rely on manually engineered features such as parse trees and Bag-of-Words (BOW) that assume segments in the same data practice category share similar syntactic structures and share similar word distributions, respectively. However, segments in the same data practice category can have diverse sentence structures, lengths, and word choice.

Given the issues with conventional ML-based analyses, scholars have begun to adopt the DL-based CNN that applies multiple layers of non-linear transformations to automatically learn features from input text without manual feature engineering (Harkous et al. 2018). While attaining superior performance over their ML counterparts, CNN-based methods only capture short sequential word dependencies and can miss long sequences of text in privacy policies. In addition, CNNs are black-box models wherein end-users cannot identify the most informative text features (via feature weights or visualizations) that led to the classifier's final decision. Taken together, these limitations necessitate an alternative DL-based approach that can capture both short and long sequential word dependencies to generate better segment representation for data practice annotation as well as output weights for input features. Therefore, we review Structured Self-Attentive Sentence Embedding (SSASE) as a nascent DL-based text classification model next.

Structured Self-Attentive Sentence Embedding (SSASE)

SSASE generates text representations for multi-class classification tasks based on a Bidirectional Long Short-Term Memory (Bi-LSTM) model with a multi-head self-attention mechanism (Lin et al. 2017). Bi-LSTM is a prevailing DL model often employed in text analytics tasks to capture sequential and contextual dependency information from text input. BiLSTMs have been shown to consistently outperform CNN-based methods in various text analytics tasks (Ebrahimi et al. Forthcoming; Samtani et al. Forthcoming). The multi-head self-attention

mechanism extracts the non-sequential global dependencies of the inputs that the conventional Bi-LSTM does not capture. Each head emphasizes different words in the text input based on one or more specific semantic aspects such as rare words and word positions (Voita et al. 2020). Figure 3 depicts the overall SSASE and its operations.

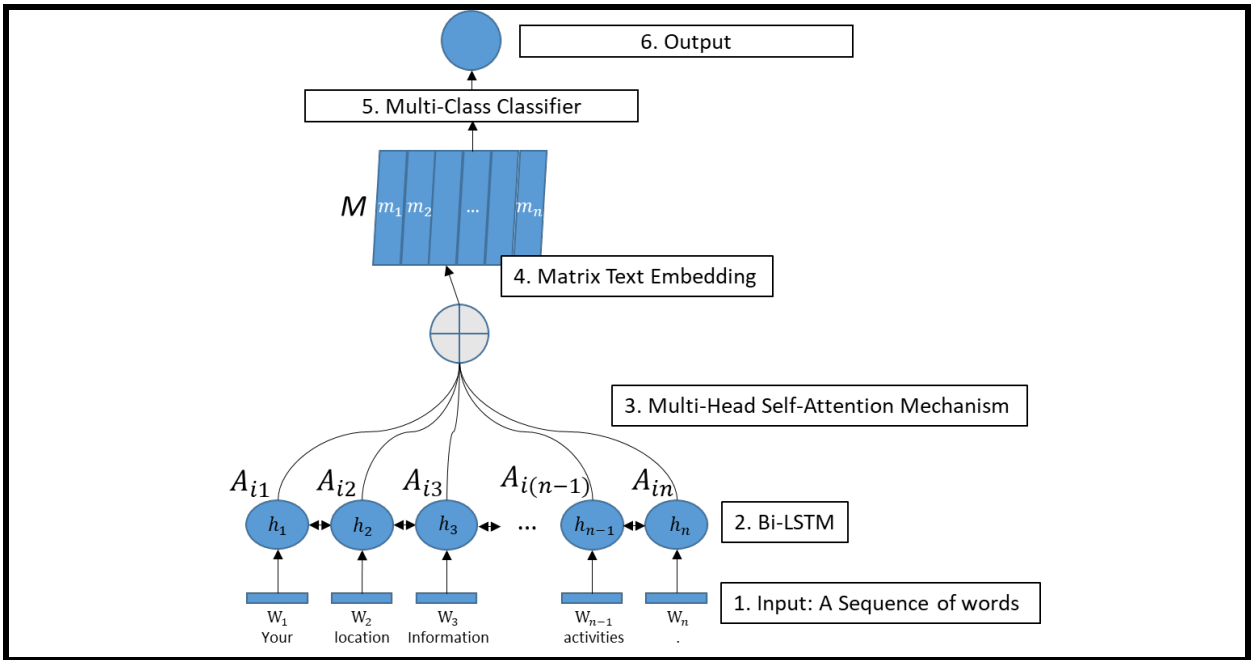


Figure 3. Conceptual Schematic of SSASE (Adapted from Lin et al. 2017)

The Bi-LSTM component of SSASE processes word sequences in both forward and backward directions. Each direction generates a "directional" hidden state $\overrightarrow{h_i}$ (or $\overleftarrow{h_i}$) based on the word embedding w_i and the previous hidden state $\overrightarrow{h_{i-1}}$ (or $\overleftarrow{h_{i+1}}$). By concatenating $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ from the forward and backward directions, $h_i = \left[\overleftarrow{h_i}\ \overrightarrow{h_i}\right]^T$ represents a more comprehensive summary of the current hidden state. The multi-head self-attention mechanism takes the Bi-LSTM hidden states $\mathbf{H} = (h_1, h_2, \ldots h_n)$ as input and learns the multi-head self-attention weight matrix A generated by: $softmax(W_{s2}\, tanh(W_{s1}H^T))$, where H^T is a transposed hidden state matrix containing semantic information captured by Bi-LSTM. W_{s1} and W_{s2} are two trainable weight matrices. W_{S1} generates linear combinations of the hidden states to capture the global correlations between terms within

the input sequence (e.g., meaningful phrases, word correspondence). W_{S2} weighs these correlations differently to extract disparate semantic aspects of the input sequence. Thus, the multi-head self-attention weight matrix A summarizes "how much attention" should be paid to each word in the input based on its contributions to the final output prediction. The weights assigned to each word in each head learned by the multi-head self-attention mechanism can be visualized to pinpoint which parts of the text are encoded in the matrix text embedding. Each column in A corresponds to one word, while each row (head) captures salient word correlations that potentially represent a latent aspect of the text. A penalty term P in the loss function is often used to diversify attention heads and avoid learning duplicated latent aspects: $\mathrm{P} = \|AA^T - I\|_F^2$, where I is the identity matrix and $\|\cdot\|_F$ is the Frobenius norm. The penalty term reaches the minimum when A is orthogonal (i.e., each head does not correlate). Matrix text embedding M is the dot product of A and H, representing important aspects of the text. M is then flattened into a vector for a classifier composed of a fully connected layer and a Softmax layer, which indicates how information extracted in each aspect can contribute to the multi-class classification.

SSASE and its variants have demonstrated their effectiveness in multi-class text classification applications, including grouping news articles (Tao et al. 2019), categorizing health records (Sousa et al. 2018), and analyzing social media sentiments (Zhao et al. 2018). However, there is limited work examining its performance in multi-label classification tasks like data practice segment annotation (needed for comparing privacy policies pre- and post-regulation implementation). In multi-label classification tasks, selected heads of the input data should be emphasized more than others. For example, FP and TP may have shared heads such as "user information." However, they also have differentiating heads that help delineate between practices. Examples include "we collect your information for …" in FP and "we share your information with

…" in TP. Improving multi-label prediction performance can be contingent upon reducing the impact of the shared heads and emphasizing key differentiating heads. Therefore, extending the multi-head attention mechanism in SSASE to differentiate head importance for effective data practice segment annotation is required. We next review attention mechanisms to ground the proposed approach for enhancing SSASE.

Attention Mechanisms

Attention mechanisms aim to identify how input features affect model performance (Du et al. 2019). Formally, attention mechanisms operate by mapping a query vector Q and a set of key vector-value vector pairs (K, V) to an output vector O (Vaswani et al. 2017). Q can be considered as a representation of focus of interest and K is a representation of the characteristics of the input. O is computed as a weighted V, where weights are alignment scores calculated based on the relationship (e.g., similarity) between Q and K. If elements in K are closely related to Q, higher alignment scores are assigned.

Attention mechanisms can be categorized into two major groups: general attention and self-attention (Du et al. 2019). The former calculates the alignment score between Q and K, and the latter calculates the alignment score within the elements in K (i.e., $Q = K$). Self-attention mechanisms have been extensively incorporated in prevailing sequence models to capture global feature dependencies for generating high-quality text representations for neural machine translation, sentiment analysis, and other applications (Letarte et al. 2018; Vaswani et al. 2017). Evaluation of attention mechanisms is typically executed by comparing the model with the attention mechanism against the one without the attention mechanism on the ground-truth dataset (Samtani et al. Forthcoming).

SSASE leverages the self-attention mechanism to capture each word's contributions to a segment's semantic. In addition, since data practice segments are usually long and cover diverse semantic aspects, single-head self-attention may miss certain aspects. SSASE leverages multi-head self-attention to jointly learned weights from different vector subspaces that focus on different semantic aspects. However, not all heads are equally important. Recent studies have shown that only a small subset of heads play important and linguistically interpretable roles in overall model performance (Voita et al. 2020). In data practice segment annotation, the importance of each head depends not only on each data practice category but also on the relationship between categories (i.e., shared and differentiating information). Existing studies considered the average or the maximum of weights over attention heads, which do not explicitly consider the varying importance of different heads in multi-label classification (Tang et al. 2019; Voita et al. 2018). Therefore, there is a need for a new approach to distinguish the importance of attention heads. However, how to adapt the self-attention mechanism to determine the importance of each attention head to produce a better representation of each data practice segment than the conventional self-attention mechanism in SSASE requires further investigation.

RESEARCH GAPS AND QUESTIONS

We identify several key research gaps within extant literature. IS scholars have made significant progress in multiple areas of information privacy research. However, the prevailing behavioral and economic methodologies adopted in prior IS literature are not intended to operate on the rich and complex text of privacy policies, which are served as the main instruments in conveying companies' data practices. As a result, little work has identified how data practices in companies' privacy policies have changed after new regulations have been implemented. Since it is critical for companies to assess regulatory compliance and for legislators to examine the impact

of privacy regulations on companies' privacy policies, there is a need for an automated approach to annotate segments in privacy policies to facilitate privacy policy evolution analysis. However, many past privacy policy analysis studies rely on manual analysis or traditional ML approaches which require manual feature engineering. Both approaches are labor-intensive and not viable for analyzing evolving application domains (e.g., privacy policy changes). While DL-based text analytics offers a promising approach for automatically extracting salient features from text data at scale, the existing methods cannot capture long sequential word dependencies, a common data characteristic in data practice segments in privacy policies. In addition, they cannot generate representations that can be visualized for potential end-users (e.g., regulators, companies, and privacy researchers) to understand the key phrases and concepts contained within each data practice category. Although emerging text analytics approaches such as SSASE can capture long sequential word dependencies and enable visualization to help stakeholders analyze privacy policies, they cannot identify key differentiating aspects for multi-label classification in analyzing privacy policy segments. Based on these research gaps, we post the following research questions for study:

- How can the nascent multi-head self-attention text analytics approach be enhanced to identify key differentiating semantic aspects in data practice segments to further improve the performance of multi-label data practice segment annotation?
- How can the enhanced automated data practice segment annotation system help analyze privacy policy evolution?

PROPOSED PRIVACY POLICY EVOLUTION ANALYTICS FRAMEWORK

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

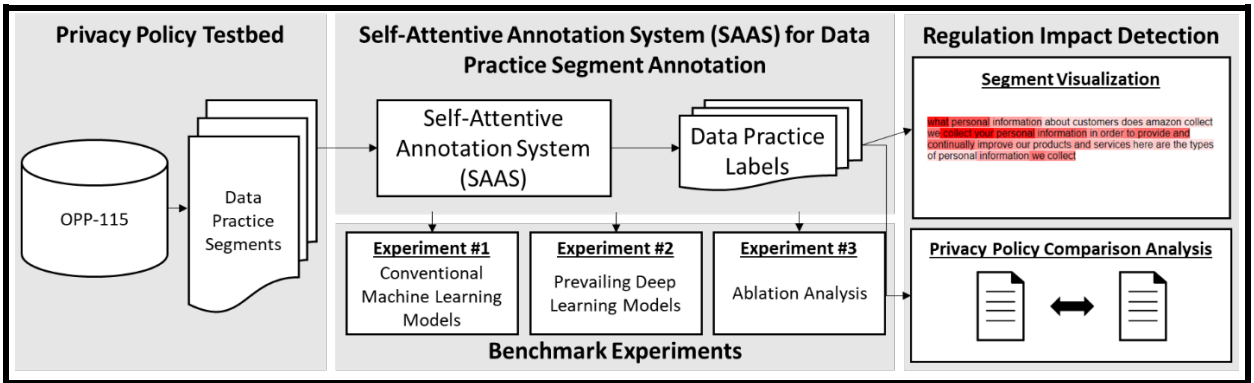


Figure 4. Proposed Privacy Policy Evolution Analytics Framework

Privacy Policy Testbed

We adopted OPP-115 (Wilson et al. 2016) as our privacy policy analysis testbed. OPP-115 is suitable for training an automated privacy policy analytics system because the legal assumptions made in the corpus annotation scheme are consistent with the transparency principle in recent privacy regulations (e.g., GDPR and CCPA) (Leone and Di Caro 2020; Poplavska et al. 2020). Although privacy policies in OPP-115 were collected before the release of several recent privacy regulations, existing data practice segment annotation systems trained on OPP-115 have been extensively used to annotate unlabeled privacy policies published after regulations (e.g., GDPR and CCPA) were enforced (Linden et al. 2020). OPP-115 contains 3,792 segments manually segmented from 115 English privacy policies of well-known websites. Each segment was

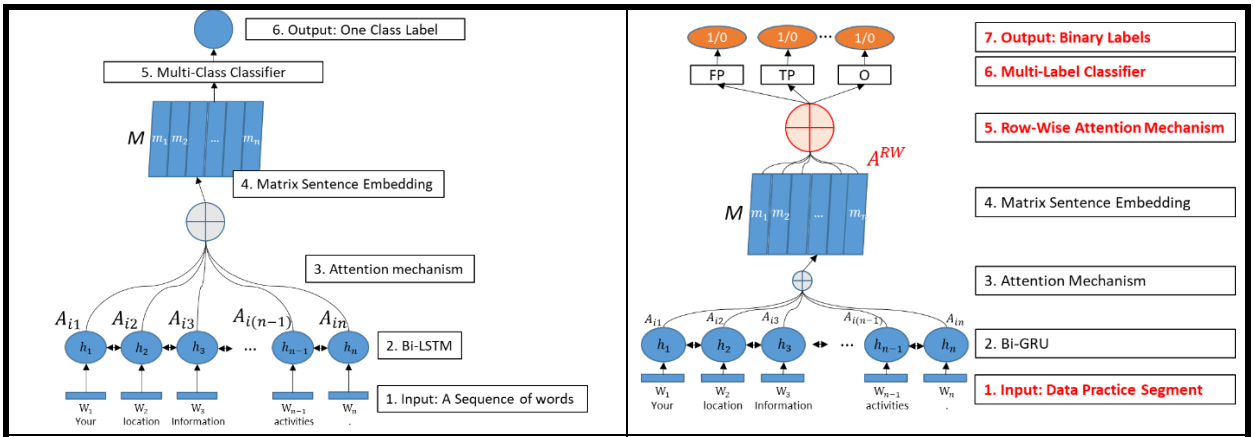
annotated with one or more data practice labels from ten categories. Consistent with previous studies, we retained data practice category labels for each segment when two or more annotators agreed on labels (Harkous et al. 2018). Forty-three segments were removed due to the lack of label agreement, resulting in 3,749 data practice segments that contain at least one data practice category label.

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

Recognizing the key limitations of the conventional multi-head self-attention embedding in the SSASE pertaining to data practice segment annotation, we propose a novel SAAS with three functions, each accounting for some essential domain requirements. First, SAAS automatically annotates privacy policy segments as one or more data practice categories (i.e., multi-label classification). This helps facilitate data practice category-level privacy policy evolution analysis. Second, the proposed SAAS can leverage differentiating attention heads to improve the performance of multi-label data practice annotation tasks. Third, it helps to visualize important text features that contribute the most to the semantics in a word sequence. Taken together, these three functions can help stakeholders more efficiently and effectively examine whether privacy policies comply with new regulations without reading large amounts of text.

SAAS is comprised of a novel Row-Wise Self-Attentive Sentence Embedding (RWSASE) model and a Multi-Label Classifier. RWSASE extends SSASE and weighs differentiating heads extracted from the text to facilitate multi-label classification. The Multi-Label Classifier classifies the learned representations into one or more data practice categories. Figure 5 compares SSASE and the proposed SAAS. Our novelties are highlighted in red. The following sub-sections further summarize each key SAAS component.

Conventional SSASE	Proposed SAAS



Conventional SSASE Procedure

1. Input: A Sequence of Words
2. Bi-LSTM
3. Attention Mechanism
4. Matrix Sentence Embedding
5. Multi-Class Classifier
6. Output: One Class Label

Proposed SAAS Procedure

Row-Wise Self-Attentive Sentence Embedding (RWSASE):

1. Input: Data Practice Segment
2. Bi-GRU
3. Attention Mechanism
4. Matrix Sentence Embedding
5. Row-Wise Attention Mechanism

Multi-Label Classifier:

6. Multi-Label Classifier
7. Output: Binary Labels

Figure 5. Illustration of Conventional SSASE and Proposed SAAS

Row-Wise Self-Attentive Sentence Embedding (RWSASE)

RWSASE builds upon the basic architecture of SSASE. It takes a data practice segment as input, where each word in the segment is represented by a word embedding. A Bi-GRU automatically extracts the forward and backward context information from the embedding sequence. Bi-GRU is a variant of Bi-LSTM, which streamlines the Bi-LSTM three-gate structure into two gates. As a result, Bi-GRU requires less parameters and can therefore converge faster than the Bi-LSTM. The multi-head self-attention mechanism learns from all hidden states the contributions of the words to the segment in every latent aspect 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 matrix sentence embedding, a low-dimensional matrix representation of the segment that separately encodes semantics of different latent aspects.

In SSASE, the downstream task that uses the matrix sentence embedding is multi-class classification. A classifier calculates the probabilities of the input segments belonging to target categories, where the contributions of latent aspects in the matrix sentence embedding are correlated. The classifier assigns lower coefficients to the shared aspects (e.g., common words), as they are not as informative as other aspects. However, the downstream task in SAAS is multi-label classification, where multiple classifiers make decisions independently for each category. With SSASE, each classifier is unaware of aspects shared among categories and may assign higher coefficients to them, resulting in misclassification. In order to suppress these shared aspects before inputting them into the classifiers, we design a novel row-wise attention mechanism (i.e., RWSASE) to emphasize differentiating aspects in the matrix sentence embedding. Inspired by self-attention mechanisms in determining the importance of each word that contributes to the text semantic, the proposed row-wise attention mechanism learns the importance of each head in multi-head attention that contributes to the segment semantic as follows:

$$A^{RW} = \text{softmax}(\tanh(W^{RW} M^{T})),$$

where A^{RW} is the row-wise attention weight vector, M^{T} is the transposed matrix sentence embedding, and W^{RW} is the trainable weight vector. Each element in A^{RW} indicates the importance of each head to the decision process. Higher weights are assigned to differentiating heads as they contribute more to delineating output labels. A^{RW} is applied back to the matrix sentence embedding by $A^{RW} * M$ for the downstream multi-label task. A^{RW} is also applied back to the multi-head attention weight matrix. The weighted multi-head attention weight matrix can be visualized to highlight the critical text features in data practice segments.

Multi-Label Classifier

A multi-label classifier contains k independent binary classifiers, corresponding to the number of output labels ($k = 10$ in our study). Each classifier takes a flattened matrix sentence embedding weighted by row-wise attention weight vector as an input. Each classifier is composed of a fully connected layer and a Sigmoid layer. The Sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ predicts the output $y_i \in \{0,1\}$, where $y_i = 1$ denotes that a segment belongs to the i-th data practice category, otherwise $y_i = 0$. The final output of SAAS is a tuple $Y = (y_1, y_2, \ldots, y_k)$ that aggregates the predictions of a data practice segment from each binary classifier.

Benchmark Experiments

Consistent with computational design science principles (Rai 2017) and DL literature, we evaluated the proposed SAAS with three benchmark experiments: (1) SAAS vs Conventional Machine Learning Models, (2) SAAS vs Prevailing Deep Learning Models, and (3) Ablation Analysis. We summarize each experiment, model category, and benchmark model in Table 5.

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-based (TF-IDF): Sentence embedding is learned based on term frequency divided by Inverse Document Frequency (IDF).	TF-IDF + LR	Mysore Sathyendra et al. 2017; Tesfay et al. 2018; Zimmeck et al. 2019
		TF-IDF + SVM	
		TF-IDF + RF	
		TF-IDF + NB	
		TF-IDF + KNN	
SAAS vs Prevailing 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.	Bi-LSTM + Max Pooling	Lin et al. 2017
		Bi-LSTM + Mean Pooling	
		Bi-GRU + Max Pooling	
		Bi-GRU + Mean Pooling	
	Attention-based: A self-attention mechanism and bi-directional recurrent structure learn a matrix sentence embedding.	SSASE	

Ablation Analysis	SAAS without the row-wise attention	Samtani et al. Forthcoming ; Zhu et al. 2021
	SAAS with the row-wise attention but without the activation function	

***Note:** Bi-: Bi-directional; CBOW: Continuous Bag-of-Words; 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: Naïve Bayes; RF: Random Forest; SVM: Support Vector Machine
******Naïve Bayes does not work with Doc2Vec text representation because it requires positive numeric input.

In Experiment 1, we compared SAAS against five conventional machine learning benchmark models commonly used in IS literature (Kitchens et al. 2018): Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbors (KNN). Two text representation techniques commonly used in privacy policy analysis literature, Doc2Vec and TF-IDF, were used to produce a text representation from each segment for input into each model (Sathyendra et al. 2017; Tesfay et al. 2018; Wilson et al. 2016; Zimmeck et al. 2019). In Experiment 2, we compared SAAS' performance against DL-based models in four categories: CNN-based, uni-directional RNN-based (LSTM and GRU), bi-directional RNN-based (Bi-LSTM and Bi-GRU), and attention-based (SSASE). Consistent with best practices in IS literature, we implemented each model used in Experiments 1 and 2 based on the details (e.g., structure, layers, parameters, etc.) provided by previous privacy policy analysis and DL literature. In Experiment 3, we conducted an ablation analysis that evaluated two variations of the SAAS: (1) SAAS without the row-wise attention to identify how the proposed row-wise attention mechanism affects the model performance and (2) SAAS with the row-wise attention but without the activation function to evaluate the effect of increasing the number of trainable parameters.

Consistent with privacy policy analysis literature, we executed each benchmark experiment with the 3,749 segments from the OPP-115 dataset that possess one or more data practice labels agreed upon by the original annotators (Wilson et al. 2016). Overall, 2,848 segments have one label, 792 segments have two labels, 88 have three labels, 18 segments have four labels and three

segments have five labels. We present the number of segments associated with each data practice category (label) in Table 6.

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***

*** Note:** The size of this testbed is consistent with or exceeds the dataset sizes used in several past IS studies that trained and evaluated DL models (Zhu et al. 2020).

The number of segments in each category ranges from 32 to 1,763, suggesting that the distribution of segments across the categories is imbalanced. The FP and TP categories have the highest number of segments with 1,522 and 1,186, respectively, while the DNT category contains the least number of segments (32). We executed each experiment for each data practice category. Since the dataset is imbalanced, model performances were measured using the precision, recall, and F1-score metrics. Precision is defined as:

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

where $c_i \in C$. C is the set of ten 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. Precision measures whether a model can correctly classify a segment into a specific category. Recall is defined as:

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

where $FN(c_i)$ (False Negative) is the number of segments incorrectly classified as not a specific data practice category c_i. Recall measures whether a model can detect all the segments in each data practice category. F1-score is the harmonic mean of precision and recall and is formulated as:

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

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). 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 on each category and across all categories. HL is a frequently used metric for evaluating the performance of algorithms operating in multi-label classification tasks (Tsoumakas and Katakis 2007). It 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 will predict data practice segments with incorrect data practice categories.

For all experiments, 5 times 2-fold (5 x 2) cross-validation was adopted (Demšar 2006; Dietterich 1998). 5 x 2 cross-validation is a suitable approach for comparing two classifiers on a single dataset. It 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 variance of performance across folds. In each fold, one partition was used for testing and the other partition was used for training (90%) and validation (10%). This process was repeated five times, and results were averaged to produce a single estimation. Paired t-tests (Demšar 2006; Dietterich 1998) were used to identify statistically significant differences between performance metrics. Performance differences were considered significant at $p < 0.05$, 0.01, and 0.001.

All experiments were executed on a single Microsoft Windows 10 Pro server with 128GB of Random Access Memory (RAM), an Nvidia GeForce GTX 1070 Ti Graphical Processing Unit (GPU), and an E5-2670 v4 at 2.60 Gigahertz (GHz) Intel Central Processing Unit (CPU). All algorithm implementations were based on the PyTorch (Paszke et al. 2019), Natural Language Toolkit (NLTK) (Bird et al. 2009), NumPy (Oliphant 2006; Van Der Walt et al. 2011), pandas (McKinney 2010), genism (Rehurek and Sojka 2010), and scikit-learn (Fabian et al. 2011) packages. To facilitate scientific reproducibility, the full SAAS implementation details are provided in Appendix A.

RESULTS AND DISCUSSION: EXPERIMENTS AND CASE STUDY

Experiment 1 Results: SAAS vs. Conventional Machine Learning Models

Experiment 1 evaluated the performance of SAAS against conventional ML models. All models were evaluated based on micro-averaged precision, micro-averaged recall, micro-averaged

F1-score, and micro-averaged HL. Results are grouped based on the underlying text representation used by the algorithm: term frequency-based or paragraph vector-based. Table 7 summarizes all model performances. The highest scores are highlighted in boldface. Precision, recall, F1-score, and HL scores for each of the ten data practice categories are reported in Appendix B.

Table 7. Experiment 1 Results: 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		0.809	**0.714**	**0.758**	**0.058**

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

SAAS outperformed all conventional ML methods, except TF-IDF + RF, on micro-averaged precision (0.809) by statistically significant margins. Moreover, SAAS achieved the best micro-averaged recall (0.714), micro-averaged F1-score (0.758), and micro-averaged HL (0.058), significantly outperforming all benchmarks. 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). 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 belonging to "Third Party Sharing/Collection," and the keyword "collect" can indicate both "First Party Collection" and "Third Party Sharing/Collection." Doc2Vec generated similar segment representations containing key terms that share similar surrounding texts (e.g., "share your information" and "collect your information"), leading to lower performances than term frequency-based models and the proposed SAAS. Furthermore, LR and SVM outperformed other term

frequency-based models on the micro-averaged F1-score, possibly due to their ability to process high dimensional features (Kamath et al. 2018). Finally, SAAS outperformed term frequency-based models on the micro-averaged F1-score.

SAAS' performance is likely attributable to its ability to leverage both context and local semantic information of each word. To illustrate SAAS's better performance, we select two example segments where SAAS correctly classified them as not belonging to FP. At the same time, 8 out of 9 benchmark approaches, including TF-IDF + LR that achieved the second-highest performance, incorrectly classified them as FP. We present the two example segments that were captured by SAAS but missed by TF-IDF + LR (the best performing benchmark model in terms of F1-Score) in Table 8. The table is organized based on the company that the privacy policy belongs to, the segment in the privacy policy, the ground truth data practice category of the segment, and the predicted data practice category generated by SAAS and TF-IDF + LR. The shades of the color encode the normalized word weights learned by each method. Dark red indicates the higher importance of phrases in the segment. Instances related to FP and TP are selected because FP and TP have the highest number of labels (653 and 548), which provide a better chance to identify the patterns that our proposed methods can capture while others cannot. A similar pattern occurs in 346 out of 1,872 segments in the testing dataset.

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

Company	Segment	Ground Truth	SAAS's Predictions	TF-IDF + LR's Predictions
Fox Sports	to deliver video content in the ordinary course of our business we may share information that we collect from or about you with third parties in order to process your requests and fulfill your orders for video content offered through the fsd services for example to process your request to view certain authenticated video content through the fsd services we may need to share information about you with your designated television content provider and or our third party service providers in order to authenticate and or authorize you to view authenticated and or localized video content through the fsd services	TP	TP	FP

Allstate Insurance	also our business partners assist us with monitoring information including but not limited to ip addresses domain names and browser data which can help us to better understand how visitors use websitesurl	TP	TP	FP

As highlighted in the yellow box in Table 8, phrases such as "collect from or about you" can indicate both "First Party Collection" and "Third Party Sharing/Collection." SAAS assigned lower weights to those phrases. Thus, decreasing the possibility of incorrect predictions misled by those phrases. In contrast, term frequency-based and paragraph vector-based methods often lack the ability to distinguish two data practice categories that share similar keyword distributions. In addition, as highlighted in the blue box in Table 8, the phrases "need to share," "third party service," and "our business partners" are more likely to indicate belonging to "Third Party Sharing/Collection," which were assigned higher weights. This example suggests that SAAS's use of the attention mechanism with Bi-LSTM enables the model to better leverage the context information to achieve higher classification performance.

Experiment 2 Results: SAAS vs. Prevailing Deep Learning Models

Experiment 2 evaluated the performance of SAAS against ten selected state-of-the-art deep learning-based models. Table 9 summarizes model performances, grouped by CNN-based, RNN-based (uni-directional and bi-directional), and attention-based models. All models were evaluated based on micro-averaged precision, micro-averaged recall, micro-averaged F1-score, and micro-averaged HL as in Experiment 1. The highest scores are highlighted in boldface. Precision, recall, F1-score, and HL of each data practice category are also reported in Appendix B.

Table 9. Experiment 2 Results: SAAS vs. Prevailing 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***
Uni-directional 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**

Bi-directional 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	**SSASE**	0.770***	0.729	0.749*	0.062***
	Proposed SAAS	**0.809**	0.714	**0.758**	**0.058**

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

As shown in Table 9, our proposed SAAS achieved the best micro-averaged precision (0.809), micro-averaged F1-score (0.758), and micro-averaged HL (0.058) at statistically significant margins. CNN and uni-directional RNN-based methods had similar performances and were both worse than those of bi-directional RNNs. This result indicates that capturing both forward and backward directions provides more comprehensive local context information for distinguishing segment semantics. Attention-based methods (SAAS and SSASE) outperformed bi-directional RNN-based methods. This finding indicates that capturing both global and local correlations with the self-attention mechanism can further disambiguate word semantics. Compared to the second-best performing model SSASE, SAAS provided additional weights for different semantic aspects in the matrix sentence embedding. As a result, it emphasized differentiating aspects that can contribute to multi-label classification and was less likely than SSASE to be misled by shared aspects (e.g., common key phrases across multiple categories).

We select two example segments (Table 10) where SAAS correctly classified them as both FP and TP. At the same time, all benchmark approaches incorrectly classified them as only FP or TP. A similar pattern occurs in 287 out of 1,872 segments in the testing dataset. We present the two example segments in Table 10. The table is organized based on the company that the privacy policy belongs to, the segment in the privacy policy, the ground truth data practice category of the segment, and the predicted data practice category generated by SAAS and SSASE. The shades of the color encode the normalized word weights learned by each method. Dark red indicates the higher importance of phrases in the segment. Instances related to FP and TP are selected because

FP and TP have the highest number of labels (653 and 548), which provide a better chance to identify the patterns that our proposed methods can capture while others cannot.

Table 10. Example Segments Detected by SAAS but Missed by SSASE

Company	Segment	Ground Truth	SAAS's Predictions	SSASE's Predictions
Fortune	these tracking technologies may be deployed 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 information collected from you through different means or at different times including both personal information and service usage information and use such information along with offline and online information obtained from other sources including third parties such as demographic information and updated contact information where that information has been lawfully disclosed to us for the purposes of learning more about you so we can provide you with relevant content	FP, TP	FP, TP	FP
Disinformation	analytic and reporting technologies like the operators of most websites we use analytic and reporting technologies to record non personal information such as internet domain and host names internet protocol ip addresses browser software operating system types clickstream patterns and the dates and times that the website and the disinformation services are accessed to do this we may use web beacons and cookies including those from third parties	FP, TP	FP, TP	FP

As highlighted in the blue box, two segments contained sentences indicating both the first party and third party will collect/access users' data. SAAS learned segment representations by considering the relationship between the given segment and all ten categories. Since FP and TP share various common semantic aspects, segments that belong to FP (or TP) are more likely to also belong to TP (or FP). In addition, SAAS can emphasize differentiating semantic aspects (e.g., "by us and or by our service providers or partners") by row-wise attention mechanism to generate representations. As a result, SAAS avoids mislabeling segments because of the increased probability caused by the shared aspects.

Experiment 3 Results: Ablation Analysis

Experiment 3 examined the effect of the row-wise attention mechanism on performance. Three variants of the row-wise attention mechanism were tested, including SAAS without the row-wise attention mechanism, SAAS with the row-wise attention mechanism but without the activation function, and SAAS with the row-wise attention. All models were evaluated based on micro-averaged precision, micro-averaged recall, micro-averaged F1-score, and micro-averaged HL as in Experiment 1. Table 11 summarizes the evaluation results. The highest scores are highlighted in boldface. Precision, recall, F1-score, and HL for each category are reported in Appendix B.

Table 11. Performance of Ablation Analysis

Model	Micro-Averaged Precision	Micro-Averaged Recall	Micro-Averaged F1-score	Micro-Averaged HL
SAAS without the row-wise attention	0.800***	0.706***	0.750***	0.060***
SAAS with the row-wise attention but without the activation function	0.802***	0.699***	0.747***	0.060***
SAAS with the row-wise attention	**0.809**	**0.714**	**0.758**	**0.058**

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

SAAS achieved the best micro-averaged precision (0.809), micro-averaged recall (0.714), micro-averaged F1-score (0.758), and micro-averaged HL (0.058) over all variants by statistically significant margins. The row-wise attention mechanism emphasized the critical semantic aspects extracted by the self-attention mechanism corresponding to each segment, thus, improving the performance. In addition, SAAS with the row-wise attention but without the activation function had a lower F1-score (0.747). The row-wise attention mechanism without the activation function is similar to add a dense layer to generate random weights for different semantic aspects learned by the self-attention mechanism. While the number of trainable parameters is the same between SAAS with the row-wise attention with and without the activation function, the results indicate that increasing the number of trainable parameters did not necessarily increase the performance. By incorporating the activation function, the row-wise attention can identify the non-linear

relationship between each word and the segment (dense layer can only identify the linear relationship), which contributes to the improvement of the performance.

Regulation Impact Detection: A Case on GDPR

To demonstrate proof-of-concept, usability, usefulness, and the potential value of our proposed SAAS, we conducted a GDPR impact detection analysis on Amazon's privacy policies. While our proposed framework can be applied for any regulation, we focus the analysis on the impacts of GDPR on privacy policy evolution. GDPR is chosen as it impacts companies worldwide 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 targeting ads without the user's free consent (violating GDPR) (Dumiak 2021). Therefore, the goal of our case study is to identify whether Amazon's pre- and post-GDPR privacy policies provide comprehensive information about the ad targeting system and if they explicitly ask users to agree to Amazon's use of their data (i.e., regulation impact detection). We employed five steps to execute the case study:

- **Step 1 – Collect privacy policies before and after a time of interest**: Since GDPR became enforceable beginning May 25, 2018, we collected the pre- and post-GDPR privacy policies released on March 3, 2014 and February 12, 2021, respectively.
- **Step 2 – Pre-process the collected privacy policies and divide them into semantically coherent segments with text segmentation techniques**: Consistent with previous studies, the retrieved privacy policies are segmented based on the HTML and <p> tags (Harkous et al. 2018). To further sub-divide long segments, we used ST-Ro (Aumiller et al. 2021), a transformer-based text segmentation model that outperforms other prevailing segmentation techniques such as GraphSeg (Harkous et al., 2018) and WikiSeg (Bertino et al. 2021).

- **Step 3 – Annotate segments using 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 FP, TP, UAED, and UCC because FP and TP contain information of users' data collected by first-party and third-party, including personally identifiable information and behavioral data that facilitate targeting ads. UAED and UCC comprise users' right to access any information collected by the first-party and third-party and opt-in/out of disclosing their data.
- **Step 5 – Visualize segments in data practice categories of interest using attention weights**: To identify the difference in privacy policies, we visualized the segments to pinpoint phrases with higher attention weight changes in each data practice category.

Table 12 presents two selected corresponding FP segments in pre- and post-GDPR privacy policies. Phrases with higher attention weights are highlighted in red.

Table 12. A Selected Segments in FP.

Time	Segment
Pre-GDPR	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 you personalized advertisements on other web sites amazon associates with content served by and web sites using checkout by amazon payment service and storage of items in your shopping cart between
Post-GDPR	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

As shown in Table 12, the segments are related to cookies, commonly known for collecting users' behavioral data to generate personalized recommendations. The pre-GDPR privacy policy states that cookies aim to provide specific features (highlighted in blue box 1) and provides several examples. In contrast, in the post-GDPR, the highlighted phrase leads users to another document

called "Cookies Notice" (highlighted in blue box 2). According to Article 25 of the EU GDPR, privacy by default states that the strictest privacy protection/settings should be applied by default without users' manual input (Information Commisioner's Office 2018). However, users will need to link to and read another document to understand Amazon's use of cookies. If users do not read "Cookies Notice," they will not know what they consent to. In this case, the post-GDPR Amazon privacy policy has increased the burden on users to understand data practices related to target ads (violating a core principle of GDPR).

DISCUSSION AND CONTRIBUTIONS

The increasing societal concern about/over consumer information privacy has led to new privacy regulations and revolutionary changes in companies' privacy policies. Consequently, there is a need to evaluate how companies change their privacy policies and whether they provide more protection to users' information. In this study, we adopted the computational design science paradigm to systematically develop a novel privacy policy evolution analytics framework. Guided by key privacy analytics domain requirements, we designed, implemented, and evaluated a novel privacy analytics framework with DL-based text analytics methods to better understand and analyze privacy policy evolution. We rigorously evaluated the proposed approach 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. As a result, our research contributes to the IS knowledge base and has managerial and practical implications. We discuss each in turn in the following sub-sections.

Contributions to the IS Knowledge Base

Novel IT artifacts often contribute prescriptive knowledge back to the IS knowledge base to guide future research (Gregor and Hevner 2013; Hevner et al. 2004; Nunamaker et al. 1990;

Peffers et al. 2007; Rai 2017; Zhu et al. 2021). Common contributions include a situated implementation of an IT artifact in a selected domain and/or design principles that can be applied other application environments. Our proposed privacy analytics framework is a situated implementation aligned at the unique intersection of information privacy and data analytics. As a result, it can potentially open a promising area of IS research inquiry in privacy analytics. It also follows three key design principles that can be applicable beyond the privacy policy evolution analysis: (1) the row-wise attention mechanism in weighing attention heads for improving the performance of multi-label classification tasks, (2) automated and quantitative measurement of privacy policy evolution analysis, and (3) representing policy changes via visualizations to reduce cognitive load. Each design principle can provide a valuable reference for future research in designing new IT artifacts for e-commerce, social media analytics, health, and privacy. Table 13 summarizes the framework components, its general design principle, the relevant IS literature to which each principle can offer value, and potential classes of research inquiry. We then further elaborate on how these design principles can offer value to each listed body of IS literature.

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

Proposed Framework Component	General Design Principle	Relevant IS Literature	Potential Class of Research Inquiry
SAAS' Attention Mechanism	The row-wise attention mechanism in weighing attention heads for improving the performance of multi-label classification tasks	E-commerce	Analyzing the relationship between consumer preference and product selection from product reviews
		Health	Identifying patients' needs for social support from social media posts
Regulation Impact Analytics	Automated and quantitative measurement of privacy policy evolution analysis	Privacy	Analysis of legal articles Analysis of Terms of Use or End-User License Agreement (EULA) documents
Visualization	Representing policy changes via visualizations to reduce cognitive load		Privacy Analytics assistance systems Privacy question/answer systems

E-commerce. Product reviews in e-commerce help to examine the relationship between users' preferences and product choices (Wu et al. 2019). However, previous studies conducted

through surveys and experiments cannot effectively analyze large amounts of unstructured data (Haws et al. 2010; Wu et al. 2019). SAAS and its extensions can facilitate an in-depth analysis of product reviews in e-commerce. For example, a product review may simultaneously complain about the texture, price, and quality of the after-sales service. The high performance of SAAS in multi-label classification can help point out multiple product issues (labels) mentioned in a product review. In addition, SAAS's ability to more effectively identify important terms or phrases in consumers' product reviews can help relevant stakeholders conduct an in-depth analysis of the cause of a complaint.

Health. Social media has become a popular channel for patients seeking health support (Bardhan et al. 2020; Chau et al. 2020). SAAS can automatically identify multiple symptoms or health needs in the content posted by patients, such as stress, physical disorders, and mental disorders. SAAS offers visualization results that can further benefit health providers in understanding how patients express their needs and how social support can satisfy different needs.

Privacy. Regulations such as GDPR and CCPA have significant and lasting global impact. The number of legal documents related to these privacy regulations exceeds what humans can manually analyze. The complex nature of legal documents makes the analysis process even more difficult. Our proposed DL-based privacy policy evolution analytics framework combined with the visualization system can help analyze a wide range of lengthy and difficult-to-understand legal articles. These includes Terms of Use, End-User License Agreement (EULA), and others. Moreover, the visualization system can play a key role in designing useful privacy analytics assistance systems.

Managerial and Practical Implications

Regulators and companies are increasingly focusing on the protection of consumer information privacy. Privacy policies are important reference documents in examining how companies handle personal data. Due to its complex characteristics, privacy policy analysis is often limited to manual analysis in extant studies. However, as businesses affected by new privacy regulations, the number of privacy policies to be analyzed and reviewed has steadily increased. Consequently, there has been a significant push in recent years to develop automated privacy policy analysis systems. However, current systems often lack accuracy, and can limit a system's decision-making capabilities. Our proposed framework helps address these limitations by enabling the analysis of hundreds or thousands of complex privacy policies in an automated fashion. We believe that our automated privacy analytics research will help the two types of stakeholders in practice: regulators and companies. We discuss each below.

Regulators. Our proposed framework can help identify content changes in different data practice categories. It can be used by regulators in different ways. First, regulators can examine whether a company's privacy policy is clear, comprehensive, and in compliance. Furthermore, when a company claims that the policy was adjusted according to the regulations, our proposed framework can be a viable and effective tool to assist in auditing the changes the company made.

Companies. Privacy policies will continue to change based on the functions provided by the business, requirements of new privacy regulations, and the evolving consumer privacy expectation. Our proposed framework can be used by companies to review potential compliance issues in their privacy policies. It can also help pinpoint and suggest specific data practice segments according to the new domestic and global privacy regulations. Consequently, companies can prevent fatal privacy violations and the associated legal and financial consequences.

CONCLUSION AND FUTURE DIRECTIONS

In this study, we developed a novel privacy policy evolution analytics framework to help identify changes in companies' privacy policies corresponding to new privacy regulations. This framework's core novelty is the proposed SAAS method. SAAS extends the nascent multi-head attention-based SSASE method to automatically label privacy policy segments that comprise multiple data practice categories. An extensive set of experiments demonstrated how SAAS outperforms conventional machine learning approaches and state-of-the-art DL and attention-based algorithms in data practice annotation. In the case study, we demonstrated the value of SAAS and the visualizations in identifying the impact of GDPR on the use of words in Amazon's privacy policies. Results of this case study indicated that instead of easing consumers' cognitive load in understanding Amazon's data practice (a key principle of GDPR), Amazon's revised privacy policy actually requires consumers to exert more effort to find all the information related to target ads. Since other geopolitical factors can impact privacy policies and new threats will continue to appear due to technological advances, privacy policies will be more frequently revised to better protect future consumer information privacy. To this end, our proposed framework addresses the increasing need for an automated privacy policy analysis tool.

There are several promising directions for future research. First, different cultures and countries may address consumer privacy issues differently. To this end, future work may regard a multilingual privacy policy evolution analytics framework to handle multiple languages in different policy data categories and segments. This can help enable international stakeholders across multiple geopolitical regions to investigate a privacy policy's global impact. Second, an AI assistive system can be developed based on SAAS to assist end-users in large-scale online privacy policy comparison from different companies when selecting online products and services.

Developing a system with a user-friendly interface and browse, search, and recommendation functions based on our proposed SAAS can help users better understand their privacy rights and data protection actions. Each direction can help build a better understanding of how organizations and consumers respond to future privacy policy requirements in a rapidly changing digital world.

REFERENCES

Acquisti, A., Brandimarte, L., and Loewenstein, G. 2015. "Privacy and Human Behavior in the Age of Information," *Science* (347:6221), pp. 509–515. (https://doi.org/10.2139/ssrn.2580411).

Adjerid, I., Peer, E., and Acquisti, A. 2018. "Beyond the Privacy Paradox: Objective versus Relative Risk in Privacy Decision Making," *MIS Quarterly* (42:2), pp. 465–488. (https://doi.org/10.25300/MISQ/2018/14316).

Akanfe, O., Valecha, R., and 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., and 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), pp. 1–21. (https://doi.org/10.1145/3403949).

Alabduljabbar, A., Abusnaina, A., Meteriz-Yıldıran, Ü., and Mohaisen, D. 2021. "TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights," in *The 20th Workshop on Privacy in the Electronic Society (WPES'21)*.

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and 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., and Xie, T. 2019. "Policylint: Investigating Internal Privacy Policy Contradictions on Google Play," in *The 28th USENIX Security Symposium*, pp. 585–602.

Anton, A. I., Earp, J. B., He, Q., Stufflebeam, W., Bolchini, D., and Jensen, C. 2004. "Financial Privacy Policies and the Need for Standardization," *IEEE Security and Privacy*, pp. 36–45. (https://doi.org/10.1109/MSECP.2004.1281243).

Antón, A. I., Earp, J. B., Vail, M. W., Jain, N., Gheen, C. M., and Frink, J. M. 2007. "HIPAA's Effect on Web Site Privacy Policies," *IEEE Security and Privacy* (5:1), pp. 45–52. (https://doi.org/10.1109/MSP.2007.7).

Aumiller, D., Almasian, S., Lackner, S., and Gertz, M. 2021. *Structural Text Segmentation of Legal Documents*. (https://doi.org/10.1145/3462757.3466085).

Bardhan, I., Chen, H., and Karahanna, E. 2020. "Connecting Systems, Data, and People: A Multidisciplinary Research Roadmap for Chronic Disease Management," *Management Information Systems Quarterly* (44:1), pp. 185–200.

Bertino, E., Kantarcioglu, M., Akcora, C. G., Samtani, S., Mittal, S., and Gupta, M. 2021. "AI for Security and Security for AI," in *CODASPY 2021 - Proceedings of the 11th ACM Conference on Data and Application Security and Privacy*. (https://doi.org/10.1145/3422337.3450357).

Bhatia, J., Evans, M. C., Wadkar, S., and Breaux, T. D. 2016. "Automated Extraction of Regulated Information Types Using Hyponymy Relations," in *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, IEEE, pp. 19–25. (https://doi.org/10.1109/REW.2016.22).

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: [Analyzing Text with the Natural Language Toolkit]*, O'Reilly Media, Inc.

Breward, M., Hassanein, K., and Head, M. 2017. "Understanding Consumers' Attitudes toward Controversial Information Technologies: A Contextualization Approach," *Information Systems Research* (28:4), pp. 760–774. (https://doi.org/10.1287/isre.2017.0706).

Buckman, J. R., Bockstedt, J. C., and Hashim, M. J. 2019. "Relative Privacy Valuations under Varying Disclosure Characteristics," *Information Systems Research* (30:2), pp. 375–388. (https://doi.org/10.1287/isre.2018.0818).

Cao, Z., Hui, K. L., and Xu, H. 2018. "An Economic Analysis of Peer Disclosure in Online Social Communities," *Information Systems Research* (29:3), pp. 546–566. (https://doi.org/10.1287/isre.2017.0744).

Cavusoglu, H., Phan, T. Q., Cavusoglu, H., and Airoldi, E. M. 2016. "Assessing the Impact of Granular Privacy Controls on Content Sharing and Disclosure on Facebook," *Information Systems Research* (27:4), pp. 848–879. (https://doi.org/10.1287/isre.2016.0672).

Chang, C., Li, H., Zhang, Y., Du, S., Cao, H., and Zhu, H. 2019. "Automated and Personalized Privacy Policy Extraction Under GDPR Consideration," in *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., and 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).

Crossler, R. E., and 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), pp. 995–1006. (https://doi.org/10.1287/isre.2019.0846).

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. 2019. "We Value Your Privacy.. Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy," *26th Annual Network and Distributed System Security Symposium (NDSS)*. (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:Jan), pp. 1–30.

Dietterich, T. G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*. (https://doi.org/10.1162/089976698300017197).

Du, M., Liu, N., and Hu, X. 2019. "Techniques for Interpretable Machine Learning," *Communications of the ACM* (63:1), pp. 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/).

Fabian, P., Michel Vincent, Olivier, G., Mathieu, B., Peter, P., Ron, W., Vanderplas Jake, and Cournapeau, D. 2011. "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* (Vol. 12).

Fawaz, K., Linden, T., and Harkous, H. 2019. "Invited Paper: The Applications of Machine

Learning in Privacy Notice and Choice," in *2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019*. (https://doi.org/10.1109/COMSNETS.2019.8711280).
Gal-Or, E., Gal-Or, R., and Penmetsa, N. 2018. "The Role of User Privacy Concerns in Shaping Competition among Platforms," *Information Systems Research* (29:3), pp. 698–722. (https://doi.org/10.1287/isre.2017.0730).
Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L. F., and Agarwal, Y. 2019. "How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices," in *12th Symposium on Usable Privacy and Security (SOUPS)*, pp. 321–340.
Gopal, R. D., Hidaji, H., Patterson, R. A., Rolland, E., and Zhdanov, D. 2018. "How Much to Share with Third Parties? User Privacy Concerns and Website Dilemmas," *MIS Quarterly* (42:1), pp. 143–164. (https://doi.org/10.25300/MISQ/2018/13839).
Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly*, pp. 337–355. (https://doi.org/10.25300/MISQ/2013/37.2.01).
Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. 2018. "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning," in *The 27th USENIX Security Symposium*, pp. 531–548.
Haws, K. L., Dholakia, U. M., and Bearden, W. O. 2010. "An Assessment of Chronic Regulatory Focus Measures," *Journal of Marketing Research* (47:5), pp. 967–982. (https://doi.org/10.1509/jmkr.47.5.967).
Heimbach, I., and Hinz, O. 2018. "The Impact of Sharing Mechanism Design on Content Sharing in Online Social Networks," *Information Systems Research* (29:3), pp. 592–611. (https://doi.org/10.1287/isre.2017.0738).
Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105. (https://doi.org/10.2307/25148625).
Information Commisioner's Office. 2018. *Guide to the General Data Protection Regulation (GDPR)*. (https://doi.org/10.1111/j.1751-1097.1994.tb09662.x).
Jain, P., Gyanchandani, M., and Khare, N. 2016. "Big Data Privacy: A Technological Perspective and Review," *Journal of Big Data* (3:1), p. 25. (https://doi.org/10.1186/s40537-016-0059-y).
Kamath, C. N., Bukhari, S. S., and Dengel, A. 2018. "Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification," in *The ACM Symposium on Document Engineering 2018 (DocEng)*, pp. 1–11. (https://doi.org/10.1145/3209280.3209526).
Kim, S. H., and Kwon, J. 2019. "How Do EHRs and a Meaningful Use Initiative Affect Breaches of Patient Information?," *Information Systems Research* (30:4), pp. 1107–1452. (https://doi.org/10.1287/isre.2019.0858).
Kitchens, B., Dobolyi, D., Li, J., and Abbasi, A. 2018. "Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data," *Journal of Management Information Systems*. (https://doi.org/10.1080/07421222.2018.1451957).
Koh, B., Raghunathan, S., and Nault, B. R. 2017. "Is Voluntary Profiling Welfare Enhancing?," *MIS Quarterly* (41:1), pp. 23–41. (https://doi.org/10.25300/misq/2017/41.1.02).
Kokolakis, S. 2017. "Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon," *Computers and Security* (64), pp. 122–134.

(https://doi.org/10.1016/j.cose.2015.07.002).
Lai, S., Xu, L., Liu, K., and Zhao, J. 2015. "Recurrent Convolutional Neural Networks for Text Classification," in *The 29th AAAI Conference on Artificial Intelligence* (Vol. 3), pp. 2267–2273.
Leone, V., and Di Caro, L. 2020. "The Role of Vocabulary Mediation to Discover and Represent Relevant Information in Privacy Policies," in *Frontiers in Artificial Intelligence and Applications*. (https://doi.org/10.3233/FAIA200851).
Letarte, G., Paradis, F., Giguère, P., and Laviolette, F. 2018. "Importance of Self-Attention for Sentiment Analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 267–275. (https://doi.org/10.18653/v1/w18-5429).
Li, X. B., and Qin, J. 2017. "Anonymizing and Sharing Medical Text Records," *Information Systems Research* (28:2), pp. 332–352. (https://doi.org/10.1287/isre.2016.0676).
Lin, Z., Feng, M., Dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. 2017. "A Structured Self-Attentive Sentence Embedding," in *The 5th International Conference on Learning Representations (ICLR)*.
Linden, T., Khandelwal, R., Harkous, H., and Fawaz, K. 2020. "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies* (2020:1), pp. 47–64. (https://doi.org/10.2478/popets-2020-0004).
McKinney, W. 2010. "Data Structures for Statistical Computing in Python," *The 9th Python in Science Conference* (445), Austin, TX, pp. 51–56. (http://conference.scipy.org/proceedings/scipy2010/mckinney.html).
Menon, S., and Sarkar, S. 2016. "Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing," *MIS Quarterly* (40:4), pp. 963–981. (https://doi.org/10.25300/misq/2016/40.4.08).
Nejad, N. M., Graux, D., and Collarana, D. 2019. "Towards Measuring Risk Factors in Privacy Policies," in *Proceedings of the Workshop on Artificial Intelligence and the Administrative State (AIAS 2019)*.
Nunamaker, J. F., Chen, M., and Purdin, T. D. 1990. "Systems Development in Information Systems Research," *Journal of Management Information Systems* (7:3), pp. 89–106. (https://doi.org/10.1080/07421222.1990.11517898).
Oliphant, T. E. 2006. "Guide to NumPy," *USA: Trelgol Publishing* (1), Trelgol Publishing USA, p. 85. (https://doi.org/10.1016/j.jmoldx.2015.02.001).
Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J., and Sadeh, N. 2018. "PrivOnto: A Semantic Framework for the Analysis of Privacy Policies," *Semantic Web* (9:2), pp. 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., and Chintala, S. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, pp. 8024–8035. (http://arxiv.org/abs/1912.01703).
Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45–77. (https://doi.org/10.2753/MIS0742-1222240302).
Poplavska, E., Norton, T. B., Wilson, S., and Sadeh, N. 2020. "From Prescription to Description:

Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme," in *Frontiers in Artificial Intelligence and Applications*. (https://doi.org/10.3233/FAIA200874).
Qamar, A., Javed, T., and Beg, M. O. 2021. "Detecting Compliance of Privacy Policies with Data Protection Laws," *ArXiv Preprint ArXiv:2102.12362.*
Rai, A. 2017. "Editor's Comments: Diversity of Design Science Research," *MIS Quarterly* (41:1), Iii–Xviii.
Rehurek, R., and Sojka, P. 2010. "Software Framework for Topic Modelling with Large Corpora," *The LREC 2010 Workshop on New Challenges for NLP Frameworks*, Citeseer, pp. 45–50.
Reidenberg, J. R., Breaux, T., Carnor, L. F., French, B., Cranor, L. F., Grannis, A., Graves, J. T., Liu, F., Mcdonald, A., Norton, T. B., Ramanath, R., Russell, N. C., Sadeh, N., and Schaub, F. 2015. "Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding," *Berkeley Technology Law Journal*.
Samtani, S., Chai, Y., and Chen, H. (Forthcoming). "Linking Exploits from the Dark Web to Known Vulnerabilities for Proactive Cyber Threat Intelligence: An Attention-Based Deep Structured Semantic Model," *MIS Quarterly*.
Samtani, S., Kantarcioglu, M., and Chen, H. 2021. "A Multi-Disciplinary Perspective for Conducting Artificial Intelligence-Enabled Privacy Analytics: Connecting Data, Algorithms, and Systems," *ACM Transactions on Management Information Systems*. (https://doi.org/10.1145/3447507).
Sathyendra, K. M., Ravichander, A., Garth Story, P., Black, A. W., Sadeh, N., and Story, G. 2017. "Helping Users Understand Privacy Notices with Automated Query Answering Functionality: An Exploratory Study," *Carnegie Mellon University Technical Report CMU-ISR-17-114R and CMU-LTI-17-005, Institute for Software Research and Language Technologies Institute, School of Computer Science*. (www.usableprivacy.org).
Sathyendra, K. M., Schaub, F., Wilson, S., and Sadeh, N. 2016. "Automatic Extraction of Opt-out Choices from Privacy Policies," in *AAAI Fall Symposium - Technical Report*.
Sathyendra, K. M., Wilson, S., Schaub, F., Zimmeck, S., and Sadeh, N. 2017. "Identifying the Provision of Choices in Privacy Policy Text," in *Conference on Empirical Methods in Natural Language Processing, Proceedings*. (https://doi.org/10.18653/v1/d17-1294).
Sechidis, K., Tsoumakas, G., and Vlahavas, I. 2011. "On the Stratification of Multi-Label Data," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. (https://doi.org/10.1007/978-3-642-23808-6_10).
Sousa, R. T., Pereira, L. A., and Soares, A. S. 2018. "Predicting Diabetes Disease Evolution Using Financial Records and Recurrent Neural Networks," *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*. (http://arxiv.org/abs/1811.09350).
Stone, E. F., Gueutal, H. G., Gardner, D. G., and McClure, S. 1983. "A Field Experiment Comparing Information-Privacy Values, Beliefs, and Attitudes across Several Types of Organizations," *Journal of Applied Psychology* (68:3), pp. 459–468. (https://doi.org/10.1037/0021-9010.68.3.459).
Story, P., Zimmeck, S., and Sadeh, N. 2018. "Which Apps Have Privacy Policies?," in *Annual Privacy Forum*, Springer, pp. 3–23. (https://doi.org/10.1007/978-3-030-02547-2_1).
Tang, G., Sennrich, R., and Nivre, J. 2019. *An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation*. (https://doi.org/10.18653/v1/w18-6304).

Tao, H., Tong, S., Zhao, H., Xu, T., Jin, B., and Liu, Q. 2019. "A Radical-Aware Attention-Based Model for Chinese Text Classification," *The 23th AAAI Conference on Artificial Intelligence* (33), pp. 5125–5132. (https://doi.org/10.1609/aaai.v33i01.33015125).

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. 2018a. "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*. (https://doi.org/10.1145/3180445.3180447).

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. 2018b. "I Read but Don't Agree: Privacy Policy Benchmarking Using Machine Learning and the EU GDPR," *The Web Conference* (2), International World Wide Web Conferences Steering Committee, pp. 163–166.

Tsoumakas, G., and Katakis, I. 2007. "Multi-Label Classification: An Overview," in *International Journal of Data Warehousing and Mining*. (https://doi.org/10.4018/jdwm.2007070101).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," in *Advances in Neural Information Processing Systems* (Vol. 2017–Decem), pp. 5999–6009.

Voita, E., Sennrich, R., Serdyukov, P., and Titov, I. 2018. "Context-Aware Neural Machine Translation Learns Anaphora Resolution," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. (https://doi.org/10.18653/v1/p18-1117).

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. 2020. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. (https://doi.org/10.18653/v1/p19-1580).

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. 2011. "The NumPy Array: A Structure for Efficient Numerical Computation," *Computing in Science and Engineering* (13:2), pp. 22–30. (https://doi.org/10.1109/MCSE.2011.37).

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., and Sadeh, N. 2016. "The Creation and Analysis of a Website Privacy Policy Corpus," in *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)* (Vol. 1), pp. 1330–1340. (https://doi.org/10.18653/v1/p16-1126).

Wu, J., Huang, L., and Zhao, J. L. 2019. "Operationalizing Regulatory Focus in the Digital Age: Evidence from an e-Commerce Context," *MIS Quarterly* (43:3), pp. 745–764. (https://doi.org/10.25300/MISQ/2019/14420).

Wunderlich, P., Veit, D. J., and Sarker, S. 2019. "Adoption of Sustainable Technologies: A Mixed-Methods Study of German Households," *MIS Quarterly* (43:2), pp. 673–691. (https://doi.org/10.25300/MISQ/2019/12112).

Zaeem, R. N., and Barber, K. S. 2017. "A Study of Web Privacy Policies across Industries," *Journal of Information Privacy and Security* (13:4), pp. 1–17. (https://doi.org/10.1080/15536548.2017.1394064).

Zaeem, R. N., and 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*. (https://doi.org/10.5220/0010180700290040).

Zhao, X., Xu, B., Zheng, D., and Li, S. 2018. "Tweet Irony Detection Using Ensembles of Word Level Attentive Long Short-Term Memory and Convolutional Neural Network," in *The 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, pp. 524–529. (https://doi.org/10.1109/FSKD.2018.8687128).

Zhu, H., Samtani, S., Brown, R., and 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), pp. 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., and Sadeh, N. 2019. "MAPS: Scaling Privacy Compliance Analysis to a Million Apps," *Proceedings on Privacy Enhancing Technologies* (2019:3), pp. 66–86. (https://doi.org/10.2478/popets-2019-0037).

Automated Analysis of Changes in Privacy Policies: A Structured Self-Attentive Sentence Embedding Approach

APPENDIX A: SAAS MODEL SPECIFICATIONS

The proposed SAAS model is implemented with PyTorch, a prevailing Python deep learning framework (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 to facilitate scientific reproducibility across varying computational setups (Zhu et al. 2021). Table A1 summarizes the specifications of each component in our proposed SAAS model, including layers, activation/dropout function implementations, and output sizes.

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)
	⋮	⋮	⋮	-	⋮
	Dense7	Dense4	Sigmoid	-	(1024, 1)

To ensure the length consistency of the input data practice segments, we padded and truncated segments that contained less than or more than (respectively) 500 words. Each word was encoded as a 300-dimensional word embedding. For the Bi-LSTM component, each direction (i.e., forward and backward) 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 1024 nodes and ten 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. We used Adam optimizer (Kingma and Ba 2015) to train the model, with a learning rate of 0.0005 and a batch size of 32. Training deep learning models with the Adam optimizer is consistent with previous IS analytics literature (Ebrahimi et al. Forthcoming).

REFERENCES

Ebrahimi, M., Nunamaker, J. F., and 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).

Kingma, D. P., and Ba, J. L. 2015. "Adam: A Method for Stochastic Optimization," in 3rd International Conference on Learning Representations, ICLR 2015.

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., and Chintala, S. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems, pp. 8024–8035. (http://arxiv.org/abs/1912.01703).

Zhu, H., Samtani, S., Brown, R., and 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), pp. 859–896. (https://doi.org/10.25300/misq/2021/15574).

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 at an aggregate level (across all data practice categories). However, we were also interested in identifying how each approach performed in each category. We present the performance of SAAS against conventional ML models by 10 data practice categories in Table B1. Models are grouped based on the underlying text representation used by the algorithm: term frequency-based or paragraph vector-based. All models were evaluated based on precision, recall, F1-score, and Hamming Loss (HL). The highest scores are highlighted in boldface.

Table B1. Performance of SAAS vs. Conventional Machine Learning 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	0.762**	0.728***	0.632***	0.565***	0.203***	0.625***	0.651***	0.406***	0.816***	0.623***
	SVM	0.784	0.757**	0.722	0.727*	0.000***	0.901	0.797***	0.559***	0.915	0.698***
	RF	0.723***	0.700***	0.620***	0.671*	0.025***	0.760*	0.887*	0.200***	0.804***	0.654***
	KNN	0.734***	0.715***	0.615***	0.562***	0.167**	0.676***	0.875**	0.617*	0.809***	0.579***
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	**0.825**	**0.853**	**0.841**	**0.873**	**0.852**	**0.950**	**0.975**	0.700	**0.953**	**0.817**
	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		0.801	0.810	0.771	0.799	0.622	0.843	0.951	0.960	0.908	0.803
Recall											
Paragraph Vector-based (Doc2Vec)	LR	0.634***	0.561***	0.356***	0.443***	0.106	0.520***	0.656*	0.535	0.731***	0.398***
	SVM	0.630***	0.532***	0.307***	0.270***	0.000**	0.397***	0.635**	0.471	0.679***	0.327***
	RF	0.544***	0.456***	0.227***	0.090***	0.002**	0.137***	0.434***	0.020***	0.251***	0.369***
	KNN	0.555***	0.479***	0.285***	0.111***	0.006**	0.262***	0.592***	0.065***	0.319***	0.539*
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	0.626***	0.600***	0.347***	0.176***	0.139	0.524***	0.416***	0.084***	0.670***	0.518***
	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		**0.848**	**0.811**	0.519	**0.642**	0.119	**0.696**	0.716	0.506	**0.821**	0.581
F1-score											
Paragraph Vector-based (Doc2Vec)	LR	0.692***	0.633***	0.454***	0.495***	0.139	0.566***	0.650***	0.455**	0.771***	0.485***
	SVM	0.698***	0.625***	0.427***	0.392***	0.000**	0.549***	0.704***	0.501*	0.779***	0.444***
	RF	0.621***	0.552***	0.331***	0.156***	0.004**	0.231***	0.580***	0.036***	0.379***	0.472***
	KNN	0.632***	0.573***	0.388***	0.184***	0.012***	0.377***	0.705***	0.116***	0.456***	0.558***
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		**0.823**	**0.808**	**0.611**	**0.709**	0.187	**0.761**	0.814	0.641	**0.862**	**0.671**
HL											
Paragraph Vector-based (Doc2Vec)	**LR**	0.197***	0.176***	0.098***	0.039***	0.033***	0.046***	0.023***	0.011***	0.035***	0.246***
	SVM	0.189***	0.174***	0.094***	0.036***	0.025*	0.038***	0.018***	0.008***	0.031***	0.239***
	RF	0.231***	0.201***	0.104***	0.042***	0.026**	0.053***	0.021***	0.008***	0.066***	0.241***
	KNN	0.225***	0.193***	0.102***	0.042***	0.025*	0.050***	0.016***	0.008***	0.062***	0.249***
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	0.210***	0.168***	0.094***	0.040***	0.023	0.038***	0.024***	0.008***	0.041***	0.206***
	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		**0.128**	**0.105**	**0.075**	0.023	0.024	**0.026**	0.011	0.004	**0.021**	**0.165**

As shown in Table B1, SAAS outperformed all conventional machine learning methods, except TF-IDF + RF, on FP (0.801), TP (0.810), UCC (0.771), PC (0.951), and O (0.803) on precision. In addition, SAAS achieved the best recall on FP (0.848), TP (0.811), UAED (0.642), DS (0.696), and ISA (0.821). Furthermore, SAAS outperformed other benchmarks on FP (0.823), TP (0.808), UCC (0.611), UAED (0.709), DS (0.761), ISA (0.862), and O (0.671) on F1-score and on FP (0.128), TP (0.105), UCC (0.075), DS (0.026), ISA (0.021), and O (0.165) on HL. The results suggest that SAAS's use of the attention mechanism with Bi-GRU enabled the model to better leverage the context information to achieve higher classification performance. However, SAAS did not outperform term frequency-based models on DR, DNT, and PC categories on all the metrics. This is mainly due to a lack of training data. These categories have less than 100 instances that the deep learning models may not extract meaningful features for classification.

We also evaluated to performances of the DL-based benchmarks by category. Table B2 summarizes model performances, grouped by CNN-based, RNN-based (uni-directional and bi-directional), and attention-based models. All models were evaluated based on precision, recall, F1-score, and HL. The highest scores are highlighted in boldface.

Table B2. Performance of SAAS vs. Prevailing Deep Learning 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											
CNN-based	**CNN**	0.809	0.783*	0.615***	0.674*	0.662	0.752**	0.795**	0.917	0.840***	0.755*
	LSTM + Max Pooling	0.823	0.823	0.614***	0.728	0.452	0.721***	0.842*	0.834*	0.860***	0.704**

Uni-directional RNN-based	**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	**SSASE**	0.797	0.795	0.662*	0.729*	0.565	0.745*	0.887	0.824	0.890	0.745
	Proposed SAAS	0.801	0.810	**0.771**	**0.799**	0.622	**0.843**	**0.951**	0.960	**0.908**	**0.803**
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	**SSASE**	0.835	0.807	0.589	0.691	0.287	0.733	0.756	0.830	0.847	0.590
	Proposed SAAS	**0.848**	**0.811**	0.519	0.643	0.119	0.696	0.716	0.506	0.821	0.581
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	**SSASE**	0.814*	0.798	0.614	0.702	0.366	0.731*	0.810	0.823	**0.867**	0.656
	Proposed SAAS	**0.823**	**0.808**	0.611	0.709	0.187	**0.761**	**0.814**	0.641	0.862	0.671
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	**SSASE**	0.133	0.111	0.084*	0.025	0.023	0.032*	0.012	**0.002**	0.021	0.180***
	Proposed SAAS	**0.128**	0.105	**0.075**	0.023	0.024	**0.026**	**0.011**	0.004	**0.021**	**0.165**

SAAS achieved the highest F1-score on FP (0.823), TP (0.808), DS (0.761), and PC (0.814) and the highest HL on FP (0.128), UCC (0.075), DS (0.026), PC (0.011), ISA (0.021), and O (0.165). SAAS achieved the best precision on UCC (0.771), UAED (0.799), DS (0.843), PC (0.951), ISA (0.908), and O (0.803). In addition, SAAS outperformed prevailing deep learning models on recall on FP (0.848) and TP (0.811). The results indicate that our proposed SAAS

generates better representations than the benchmark methods as it jointly considering unique differentiating aspects of all data practice categories. Given a segment, SAAS is less likely to mis-label the corresponding categories.

We were also interested in examining whether how the proposed SAAS without row-wise attention and with a row-wise attention but no activation function performed on each category. Table B3 summarizes model performances of SAAS and its two variants by category. All models were evaluated based on precision, recall, F1-score, and HL. The highest scores are highlighted 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										
SAAS without the row-wise attention	0.787	0.811	0.758	**0.800**	**0.748**	**0.861**	0.924	0.930	0.876*	**0.825**
SAAS with the row-wise attention but without the activation function	0.785	**0.818**	0.751	0.780	0.745	0.821	0.928	**0.970**	**0.916**	0.816
SAAS	**0.801**	0.810	**0.771**	0.799	0.622	0.843	**0.951**	0.960	0.908	0.803
Recall										
SAAS without the row-wise attention	0.835	0.789	**0.550**	**0.684**	**0.155**	0.676	0.727	**0.736**	**0.838**	0.555
SAAS with the row-wise attention but without the activation function	0.840	0.779	0.542	0.610	0.111	**0.706**	**0.755**	0.671	0.786	0.551*
SAAS	**0.848**	**0.811**	0.519	0.642	0.119	0.696	0.716	0.506	0.821	**0.581**
F1-score										
SAAS without the row-wise attention	0.806**	0.796	**0.629**	**0.723**	**0.240**	0.751	0.809	**0.807**	0.854	0.656
SAAS with the row-wise attention but without the activation function	0.810	0.796	0.625	0.655	0.185	0.753	**0.830**	0.783	0.842	0.655
SAAS	**0.823**	**0.808**	0.611	0.709	0.187	**0.761**	0.814	0.641	**0.862**	**0.671**
HL										

SAAS without the row-wise attention	0.140*	0.109	**0.073**	**0.022**	**0.023**	**0.026**	0.011	**0.003**	0.023	0.169
SAAS with the row-wise attention but without the activation function	0.138	0.108	0.074	0.028	**0.023**	0.027	**0.010**	**0.003***	0.024	0.169
SAAS	**0.128**	**0.105**	0.075	0.023	0.024	0.026	0.011	0.004	**0.021**	**0.165**

As shown in Table B3, SAAS outperformed its variants on the majority of data practice categories on F1-score (FP: 0.823; TP: 0.808; DS: 0.761; ISA: 0.862; O: 0.761). This is mainly because the row-wise attention can emphasize the critical semantic heads in segment embedding extracted by the multi-head self-attention mechanism. In addition, compared to SAASE with the row-wise attention but without the activation function, the results indicate that the proposed row-wise attention operation contributes to performance improvement while increasing the number of trainable parameters does not necessarily improve the performance.

APPENDIX C: SENSITIVITY ANALYSIS OF OUR PROPOSED SAAS

We assessed the effect of parameter variations on our proposed SAAS model. We examined SAAS's sensitivity to four key sets of deep learning specific 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 implementation had 256 Bi-LSTM hidden states. It also contained 256 attention units to extract aspects of segments into a 30-head matrix segment embedding and 1024 units in the dense layer. For each SAAS variation, we changed the target parameter and fixed all other parameters. We present the best performance of each parameter for each metric in bold. Table C1 summarizes the performances of the SAAS model and its variants. For each set of parameters, all models were compared with the baseline model to examine the statistical significance. The null hypothesis assumes that there is no significant difference between each model and the baseline model.

Table C1. Performance of SAAS with Parameter Variations

Model	Micro-averaged Precision	Micro-averaged Recall	Micro-averaged F1-score	Micro-averaged HL
Number of Hidden States in Bi-LSTM (Baseline Model: 256 Hidden States)				
128 Hidden States	0.806	0.714	0.757	**0.058**
256 Hidden States	**0.809**	0.716	**0.759**	**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	**0.809**	**0.716**	**0.759**	**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	**0.809**	**0.716**	**0.759**	**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	**0.809**	**0.716**	**0.759**	**0.058**
2048 Units	0.804	0.711	0.755	0.059

*: 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 0.809), micro-averaged recall (between 0.714 and 0.716), micro-averaged F1-score (between 0.757 and 0.759), and 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 and micro-averaged HL. Similarly, altering the number of attention units, attention heads, and units in the dense layer did not affect any of the performance metrics with statistically significant margins. This may suggest that SAAS model performance was not sensitive to parameter changes on the attention-unit, attention-head, and unit-in-the-dense-layer implementations. 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 to correctly classify the segment.