Automatic Detection of Vague Words and Sentences in Privacy Policies

Logan Lebanoff and **Fei Liu**
Department of Computer Science
University of Central Florida, Orlando, FL 32816, USA
loganlebanoff@knights.ucf.edu feiliu@cs.ucf.edu

Abstract

Website privacy policies represent the single most important source of information for users to gauge how their personal data are collected, used and shared by companies. However, privacy policies are often vague and people struggle to understand the content. Their opaqueness poses a significant challenge to both users and policy regulators. In this paper, we seek to identify vague content in privacy policies. We construct the first corpus of human-annotated vague words and sentences and present empirical studies on automatic vagueness detection. In particular, we investigate context-aware and context-agnostic models for predicting vague words, and explore auxiliary-classifier generative adversarial networks for characterizing sentence vagueness. Our experimental results demonstrate the effectiveness of proposed approaches. Finally, we provide suggestions for resolving vagueness and improving the usability of privacy policies.

1 Introduction

Website privacy policies are difficult to read and people struggle to understand the content. Recent studies (Sadeh et al., 2013) have raised concerns over their opaqueness, which poses a considerable challenge to both Internet users and policy regulators. Nowadays, consumers supply their personal information to online websites in exchange for personalized services; they are surrounded by smart gadgets such as voice assistants and surveillance cameras, which constantly monitor their activities in the home and work environments. Without clearly specifying how users' information will be collected, used and shared, there is a substantial risk of information misuse, including undesired advertisements and privacy breaches. Especially with recent high-profile cases involving Facebook and Cambridge Analytica, the public is becoming more aware and concerned with how their information is handled.

S1	We *may* use the *information* automatically collected from your computer or *other devices* for the following uses... (Vagueness: 3.8)
S2	In addition, in *some* cases the Sites can deliver content based on your current location if you choose to enable that feature. (Vagueness: 2.25)
S3	Our Sites and Services *may*, *from time to time*, provide links to sites operated by *third parties*. (Vagueness: 3.2)
S4	To customize and serve advertising and *other* marketing communications that *may* be visible to you on our Sites and Services or *elsewhere* on the internet. (Vagueness: 4)
S5	This includes your credit card number, income level, or *any other* information that would *normally* be considered confidential. (Vagueness: 3)

Table 1: Example human-annotated vague words and sentences. Vague words are *italicized*. Averaged sentence vagueness is given in the parentheses. Higher score is more vague.

Privacy policies are binding agreements between companies and users that stipulate how companies collect, use, and share users' personal information. They are lengthy and difficult to read. Bhatia et al. (2016) suggested two possible causes for this. First, privacy policies must be *comprehensive* in order to cover a variety of uses (e.g., in-store and online purchases). Second, the policies have to be *accurate* to all data practices and systems. Clearly, it would be difficult for a company's legal counsel to anticipate all future needs. They need to resort to vague language to describe the content, causing it to be difficult to read and compromising the effectiveness of privacy policies.

In this paper, we present the first study on automatic detection of vague content in website privacy policies. We construct a sizable corpus containing word- and sentence-level human annotations of vagueness for privacy policy documents. The corpus contains a total of 133K words and

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3508–3517
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

4.5K sentences. Our methods for automatically detecting vague words and sentences are based on deep neural networks, which have demonstrated impressive recent success. Specifically, we investigate context-aware and context-agnostic models for predicting word vagueness, where feature representations of words are built with and without considering their surrounding words. By this, we seek to verify the hypothesis that vagueness is an intrinsic property of words and has little to do with context. To understand sentence vagueness, we explore auxiliary-classifier generative adversarial networks (AC-GAN, Odena et al., 2018). The model has performed strongly on vision tasks (e.g., image synthesis), however, whether it can be adapted to handle text data has not been thoroughly investigated. We train the AC-GAN model to discriminate between real/fake privacy policy sentences while simultaneously classifying sentences exhibiting different levels of vagueness, including "clear," "somewhat clear," "vague," and "extremely vague," thus improving the model's generalization capabilities. The detected vague words and sentences can assist users in browsing privacy policy documents, and privacy regulators in assessing the clarity of privacy policy practices. Our research contributions include the following:

- we present the first study on automatic detection of vague content in privacy policies. Vague content compromises the usability of privacy policies and there is an urgent need to identify and resolve vagueness;
- we construct a sizable text corpus including human annotations for 133K words and 4.5K sentences of privacy policy texts. The data[1] is available publicly to advance research on language vagueness; and
- we investigate both context-aware and context-agnostic methods for predicting vague words. We also explore the auxiliary-classifier generative adversarial networks for characterizing sentence vagueness. This is the first study leveraging deep neural networks for detecting vague content in privacy policies.

2 Related Work

Privacy policies are often verbose, difficult to read, and perceived as ineffective (McDonald and Cranor, 2008). In particular, vague language in these documents hurts understanding. *"A term is regarded as vague if it admits borderline cases, where speakers are reluctant to say either the term definitely applies or definitely does not apply,"* a definition of vagueness quoted from (van Deemter, 2010). Legal scholars and language philosophers strive to understand vagueness from a theoretical perspective (Keefe, 2000; Shapiro, 2006). The "sorites paradox" describes the phenomenon of vagueness (Keefe, 2000). It states that small changes in the object do not affect the applicability of a vague term. For example, a room can remain "bright" even if the light is dimmed little by little until it is entirely extinguished, thus creating a paradox. Hyde (2014) further suggests that vagueness is a feature pertaining to multiple syntactic categories. Nouns, adjectives and adverbs (e.g., "child", "tall", "many") are all susceptible to reasoning. These studies often focus on linguistic case studies but not on developing resources for automatic detection of vagueness.

Recent years have seen a growing interest in using natural language processing techniques to improve the effectiveness of website privacy policies. Sadeh et al. (2013) describe a Usable Privacy Policy Project that seeks to semi-automate the extraction of salient details from privacy policies. Other studies include crowdsourcing privacy policy annotations and categorizing data practices (Ammar et al., 2012; Massey et al., 2013; Wilson et al., 2016b,a), grouping text segments related to certain policy issues (Liu et al., 2014; Ramanath et al., 2014), summarizing terms of services (Braun et al., 2017), identifying user opt-out choices (Sathyendra et al., 2017), and many others. These studies emphasize the "too long to read" issue of privacy policies but leave behind the "difficult to understand" aspect, such as identifying and eliminating vague content.

The work of (Liu et al., 2016) is close to ours. The authors attempt to learn vector representations of words in privacy policies using deep neural networks, where the vectors encode not only semantic/syntactic aspects but also vagueness of words. The model is later fed to an interactive visualization tool (Strobelt et al., 2016) to test its ability to discover related vague terms. While promising, their approach is not fully automatic, and the feasibility of detecting vague words and sentences in an automatic manner is still left untested.

In this work we conduct the first study to auto-

[1] https://loganlebanoff.github.io/data/vagueness_data.tar.gz

matically detect vague content from privacy policies. We ask human annotators to label vague words and sentences and train supervised classifiers to do the same. Classifying vague words is a challenging task, because vagueness is an understudied property and it spans multiple syntactic categories (e.g., "usually," "personal data," "necessary"). Neural network classifiers such as CNN and LSTM have demonstrated prior success on text classification tasks (Zhang and Wallace, 2015), but whether they can be utilized to identify vague terms is not well understood.

For sentence classification, we investigate auxiliary classifier generative adversarial networks (AC-GAN, Odena et al., 2018). GANs have seen growing popularity in recent years (Mirza and Osindero, 2014; Yu et al., 2016; Li et al., 2017; Gu et al., 2018; Cai and Wang, 2018). AC-GAN is a variant of GAN that generates word sequences using class-conditional probabilities. E.g., it generates "fake" privacy policy sentences exhibiting different degrees of vagueness (e.g., "clear," "vague," "extremely vague"). AC-GAN nicely combines real (human-annotated) and fake (synthetic) privacy policy sentences in a discriminative framework to improve the model's generalization capabilities. This can be equated to a semi-supervised learning paradigm through augmentation of the dataset with generated sentences. Data augmentation is particularly valuable for vagueness detection, which generally has small expensive datasets. We perform a full analysis on AC-GAN and compare it to state-of-the-art systems.

3 The Corpus

Annotating vague words and sentences is a non-trivial task. We describe our effort to select privacy policy sentences for annotation, recruit qualified workers, and design annotation guidelines.

We select 100 website privacy policies from the collection gathered by Liu et al. (2014). The documents are quite lengthy, containing on average 2.3K words. More importantly, most content is not vague. To obtain a more balanced corpus, a filtering step is used to select only sentences that have a moderate-to-high chance of containing vague content. Fortunately, Bhatia et al. (2016) provide a list of 40 cue words for vagueness, manually compiled by policy experts. We therefore retain only sentences containing one of the cue words for further annotation. A brief examination shows that most of the sentences removed from the corpus are indeed clear. Even with this bias, the resulting corpus still contains a small portion of clear sentences (See Figure 1). The reason is that a cue word can be used in a way that is not vague. For example, in the sentence "Users *may* post to our website," the word *may* indicates permission but not possibility, and therefore the sentence is not vague.

Vague Term	Freq.	Vague Term	Freq.
may	1,575	other information	30
personal info.	465	non-personal info.	30
information	302	sometimes	27
other	261	reasonably	26
some	214	appropriate	25
certain	205	necessary	24
third parties	183	certain info.	23
third party	134	typically	22
pers. iden. info.	88	affiliates	21
time to time	75	reasonable	20
most	54	non-personal	19
generally	52	personally iden.	18
personal data	52	such as	18
third-party	49	usually	17
others	41	personal	16
general	39	may be	15
many	37	content	14
various	36	otherwise	14
might	35	periodically	14
services	33	similar	14

Table 2: The most frequent vague terms identified by human annotators and their frequencies in our corpus. "pers.," "iden." and "info." are shorthand for "personally," "identifiable" and "information."

Reidenberg et al. (2015) discuss attempts to use crowdsourced workers as a cost-effective alternative to policy experts for annotating privacy policies. In this study, we hire crowd workers from the Amazon Mechanical Turk platform. To recruit quality workers, we require them to reside in the U.S. and be proficient in English; they are skilled workers maintaining a task success rate of 90% or above. We provide example labelled vague terms obtained from the case studies described in Bhatia et al. (2016) to reduce discrepancies among workers. The annotators are then asked to use their best judgment to perform the task.

Given a privacy policy sentence, the annotators are instructed to identify all vague terms[2] and assign a score of vagueness to the sentence. A vague term is limited to be 5 words or less (e.g., "including but not limited to"). We use this rule to prevent annotators from tagging an entire sentence/clause as vague. A slider is provided in the interface to al-

[2] We use "term" to denote either a single word or a phrase.

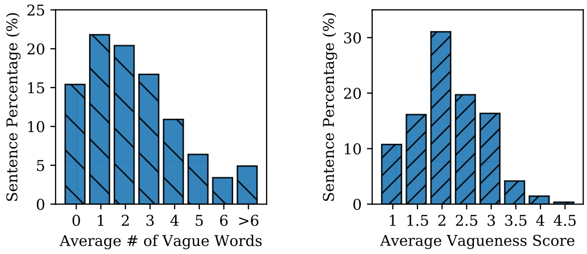

Figure 1: (Left) Percentage of sentences containing different numbers of vague words. (Right) Perc. of sentences with different levels of vagueness. 1 is clear, 5 is extremely vague.

low annotators to select a vagueness score for the sentence: 1 is extremely clear and 5 is extremely vague. We design a human intelligence task (HIT) to include 5 privacy policy sentences and a worker is rewarded $0.05 for completing the task. Five human workers are recruited to perform each task.

We obtain annotations for 133K words and 4.5K sentences. The average sentence vagueness score is 2.4±0.9. As of inter-annotator agreement[3], we find that 47.2% of the sentences have their vagueness scores agreed by 3 or more annotators; 12.5% of the sentence vagueness scores are agreed by 4 or more annotators. Furthermore, the annotators are not required to select vague words if they believe the sentences are clear. We remove vague words selected by a single annotator. Among the rest, 46.1% of the words are selected by 3 or more annotators; 18.5% of the words are selected by 4 or more annotators. These results suggest that, although annotating vague terms and sentences is considered challenging, our annotators can reach a reasonable degree of agreement. We present example vague terms in Table 2. Note that we obtain a total of 1,124 unique vague terms, which go well beyond the 40 cue words used for sentence preselection. Figure 1 shows more statistics on sentence vagueness, including (i) the percentages of sentences containing different numbers of vague words, and (ii) the percentages of sentences whose vagueness scores fall in different ranges.

4 Word Vagueness

We seek to test an important hypothesis related to word vagueness. We conjecture that vagueness is an intrinsic property of words; whether a word is vague or not has little to do with its context words. To verify this hypothesis, we build context-aware and context-agnostic models to classify each word in a privacy policy sentence as either vague or non-vague. The ground-truth labels are obtained by consolidating human annotations (see Table 3 for an example). A word is labelled 1 if it is selected by two or more annotators, otherwise 0. We describe details of the two classifiers below.

Sent: This includes your credit card number , income level , or any other information that would normally be considered confidential .

Annotator 1: any, other, normally
Annotator 2: any other information
Annotator 3: normally, confidential, any other

Ground Truth Labels: [This]$_0$ [includes]$_0$ [your]$_0$ [credit]$_0$ [card]$_0$ [number]$_0$, [income]$_0$ [level]$_0$, [or]$_0$ ***[any]***$_1$ ***[other]***$_1$ [information]$_0$ [that]$_0$ [would]$_0$ ***[normally]***$_1$ [be]$_0$ [considered]$_0$ [confidential]$_0$.

Table 3: Ground truth labels are obtained by consolidating human-annotated vague terms; "any," "other," "normally" are labelled 1 because they are selected by 2 or more annotators.

Context-aware classifier. It builds feature representations of words based on the surrounding context words. Given its strong performance, we construct a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) for this purpose. A word is replaced by its word2vec embedding (Mikolov et al., 2013) before it is fed to the model. For each time step, we concatenate the hidden states obtained from the forward and backward passes and use it as input to a feedforward layer with sigmoid activation to predict if a word is vague or non-vague. Because single words consist of the majority of the human-annotated vague terms, we choose to use binary word labels instead of a BIO scheme (Chiu and Nichols, 2016) for sequence tagging. Figure 2 shows the architecture.

Context-agnostic classifier. It uses intrinsic feature representations of words without considering the context. Specifically, we represent a word using its word2vec embedding, then feed it to a feedforward layer with sigmoid activation to obtain the prediction (Figure 2). We train the classifier using a list of unique words obtained from the training data; a word is considered positive if it has a ground truth label of 1 in any sentence, otherwise negative. Note that the ratio of positive/negative unique words in our corpus is 1068/3176=0.34. At test time, we apply the binary classifier to each word of the test set. A word is assigned the same label regardless of which sentence it appears in. We adopt this setting to ensure the context-aware and context-agnostic results are comparable.

5 Sentence Vagueness

We next investigate how vagueness is manifested in privacy policy sentences. Our goal is to assign a

[3] We choose not to calculate a kappa statistic, as labelling vague words/sentences is not a clear-cut classification task.

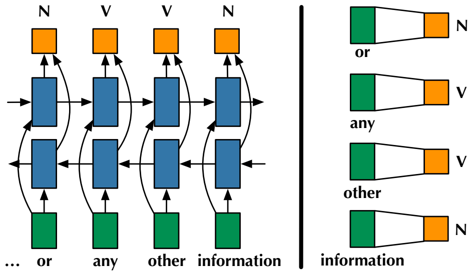

Figure 2: (Left) Context-aware word classifier implemented as a bidirectional LSTM. (Right) Context-agnostic classifier. "V" and "N" are shorthands for "vague" and "non-vague."

label to each sentence indicating its level of vagueness. We derive ground truth sentence labels by averaging over vagueness scores assigned by human annotators, and further discretizing the scores into four buckets: [1,2), [2,3), [3,4), [4,5], respectively corresponding to "clear," "somewhat clear," "vague," and "extremely vague" categories. The sentences in the four buckets respectively consist of 26.9%, 50.8%, 20.5%, and 1.8% of the total annotated sentences. We choose to predict discrete labels instead of continuous scores because labels are more informative to human readers. E.g., a label of "extremely vague" is more likely to trigger user alerts than a score of 4.2.

5.1 Auxiliary-Classifer GAN

Predicting vague sentences is a nontrivial task due to the complexity and richness of natural language. We propose to tackle this problem by exploring the auxiliary classifier generative adversarial networks (AC-GAN, Odena et al., 2018). We choose GAN because of its ability to combine text generation and classification in a unified framework (Yu et al., 2016; Li et al., 2017; Gu et al., 2018). Privacy policy sentences are particularly suited for text generation because the policy language is restricted and a text generator can effectively learn the patterns. AC-GAN has a great potential to make use of both human-annotated data and "fake" augmented data for classification. The system architecture is presented in Figure 3. The generator learns to generate "fake" privacy policy sentences and sentences exhibiting different levels of vagueness using class conditional probabilities (hence the name auxiliary-classifer GAN). The discriminator learns to discriminate among real/fake sentences as well as sentences of different levels of vagueness. They are jointly trained using a heuristic, non-saturating game loss. In the following we present the model details.

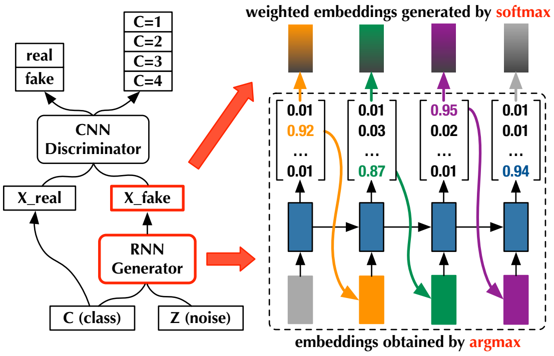

Figure 3: System architecture for AC-GAN. (Left) The Generator generates plausible privacy policy sentences (X_fake). The Discriminator must learn to differentiate between real and fake sentences as well as predicting the vagueness category (C) of the sentences. (Right) RNN-generator. A vocabulary distribution is produced for each step. Gumbel-softmax is applied to the distributions to calculate weighted embeddings to be used by the Discriminator (arrows pointing up). Argmax is applied to the distributions to retrieve embeddings to be passed to the next step (arrows pointing down).

5.2 Sentence Generator

The generator focuses on generating "fake" samples that resemble privacy policy sentences of a given vagueness category. This is denoted by $P(X|C)$, where $X = \{x_t\}_{t=1}^{T}$ is a sequence of words and $C \in \{1,2,3,4\}$ is a vagueness category. A vagueness category is randomly sampled in the generation process, and the generator attempts to generate a sentence of that vagueness level. A typical RNN text generator unrolls the sequence X one word at a time until an end-of-sentence symbol (EOS) is reached. At time step t, it samples a word x_t from a vocabulary-sized vector of probability estimates $P(x_t)$:

$$x_t \sim P(x_t) = \text{softmax}(\mathbf{a}_t), \tag{1}$$

$$\mathbf{a}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b}, \tag{2}$$

$$\mathbf{h}_t = f_{RNN}(\mathbf{h}_{t-1}, x_{t-1}), \tag{3}$$

where $\mathbf{a}_t$ is a vector of activation values and $\mathbf{h}_t$ is the t-th RNN hidden state. We train a neural text generator, implemented as Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997), on a large collection of privacy policy sentences using cross-entropy loss. While generating natural language sentences is successfully tackled by recurrent neural networks, the generated sentences are not necessarily vague. Training the generator only on vague sentences is impractical because there is a limited number of annotated sentences. In this paper we introduce a new way of defining class conditional probabilities:

$$x_t \sim P(x_t|C) = \text{softmax}(\mathbf{a}_t + \lambda_C \mathbf{v}), \quad (4)$$

where $\mathbf{v}$ is a vocabulary-sized, trainable vector indicating how likely a vocabulary word is vague. λ_C is a coefficient for vagueness category C. The underlying assumption is that a "clear" sentence is less likely to contain vague words (λ_C is negative), whereas an "extremely vague" sentence tends to contain many vague words (λ_C is positive).

Finally, the generated "fake" sentences, together with "real" sentences labelled by human annotators, are fed to the discriminator for training a classifier discriminating between real/fake sentences and sentences of different vagueness levels. Nevertheless, there remains a critical issue with the current system: we cannot backpropagate through discrete samples X. As a result, the generator parameters cannot be properly updated using backpropagation. To circumvent this issue, we attempt the reparameterization trick with Gumbel-Softmax relaxation (Gu et al., 2018).

Straight-Through Gumbel-Softmax. Two competing issues exist in the RNN generator. First, the discriminator requires a continuous form for each generated word to keep the entire model differentiable. Second, the generator requires a discrete choice for each word to generate a sentence, rather than propagating "partial words" through the sequence. To solve this problem, the *softmax* distribution over vocabulary words is sent to the discriminator, while the *argmax* over the distribution is sent to the next step of the generator. This system is referred to as Straight-Through Gumbel.

We explain the process of calculating the softmax distribution to send to the discriminator. To simulate the random-sampling process, the approach applies reparameterization to shift randomness from sampling a discrete variable x_t (Eq. (4)) to sampling a continuous noise vector $\mathbf{z}_t$ following the Gumbel distribution (Eq. (5)). The noise vector is added to the activation $\mathbf{a}_t + \lambda_C \mathbf{v}$ to compute the *argmax* (Eq. (6)). To simulate the argmax operation, a temperature parameter τ is applied to *softmax* (Eq. (7)), where small values of τ greatly skew the distribution, causing it to peak at the largest value, while still remaining differentiable. Similar reparameterization is also used for variational auto-encoders (Kingma and Welling, 2014).

$$\mathbf{z}_t \sim \text{Gumbel}(z) \quad (5)$$

$$x_t = \text{argmax}(\mathbf{a}_t + \mathbf{z}_t + \lambda_C \mathbf{v}) \quad (6)$$

$$P(x_t|C) = \text{softmax}(\frac{\mathbf{a}_t + \mathbf{z}_t + \lambda_C \mathbf{v}}{\tau}) \quad (7)$$

The generator requires a discrete word to propagate to the next time step of the RNN. The word with the maximum activation value is chosen as shown in (Eq. (6)). An illustration of ST Gumbel is presented in Figure 3.

5.3 Sentence Discriminator

A sentence discriminator learns to perform two tasks simultaneously. Given a privacy policy sentence X, it predicts a probability distribution over its sources, denoted by $P(S|X)$, where S = {real, fake}; and a probability distribution over its level of vagueness, denoted by $P(C|X)$, C = {clear, somewhat clear, vague, extremely vague}. The learning objective for the discriminator is to maximize the log-likelihood of making correct predictions on both tasks, denoted by $L_C + L_S$, where L_C and L_S are defined in Eq. (8) and (9).

$$L_C = \mathbb{E}[\log P(C = c|X_{real+fake})] \quad (8)$$

$$L_S = \mathbb{E}[\log P(S = real|X_{real})] + \mathbb{E}[\log P(S = fake|X_{fake})] \quad (9)$$

The ground truth vagueness labels C for real sentences are annotated by human annotators. For fake sentences the labels are randomly sampled in the generation process; and conditioned on the sampled vagueness labels, fake sentences are generated using $P(x_t|C)$ (Eq. (7)).

$$L'_C = \mathbb{E}[\log P(C = c|X_{fake})] \quad (10)$$

$$L'_S = \mathbb{E}[\log P(S = real|X_{fake})] \quad (11)$$

The *generator* is trained to maximize $L'_C + L'_S$ as illustrated in Eq. (10-11). Intuitively, the generator is rewarded (or punished) only based on the "fake" samples it produces. It is rewarded by generating sentences correctly exhibiting different levels of vagueness, denoted by (L'_C). It is also rewarded by generating sentences that look "real" and cannot be easily distinguished by the discriminator (L'_S). Eq. (11) corresponds to a heuristic, non-saturating game loss that mitigates gradient saturation (Goodfellow, 2016).

We experiment with two variants of the discriminator, implemented respectively using the convolutional neural networks (CNN) (Zhang and Wallace, 2015) and LSTM (Hochreiter and Schmid-

System	Word-Level P (%)	R (%)	F (%)
Context-Agnostic	11.30	**78.15**	19.71
Context-Aware	**68.39**	53.57	**60.08**

Table 4: Results of detecting vague words in privacy policies using context-aware and context-agnostic classifiers.

huber, 1997). In both cases, the discriminator assigns a source and a vagueness label to each sentence. The CNN discriminator scans through each sentence using using a sliding window and apply a number of filters to each window. A max pooling over the sequence is performed to create a feature map for the sentence. This feature map is treated as the sentence representation. It is fed to two separate dense layers with softmax activation to predict $P(C|X)$ and $P(S|X)$ respectively. In contrast, the LSTM discriminator runs a forward pass through the sentence and uses the last hidden state as the sentence representation. Similarly, this representation is fed to two dense layers used to predict $P(C|X)$ and $P(S|X)$. Both methods produce probability estimations using a shared sentence representation. Given the scarcity of labelled sentences, this multitask setting is expected to improve the model's generalization capabilities.

6 Experiments

We conduct experiments on the annotated corpus using a 5-fold cross validation; 10% of the training data in each fold are reserved for validation. In the following sections we present details of experimental settings and report results on detecting vague words and sentences in privacy policy texts.

6.1 Parameter Settings

The Xavier scheme (Glorot and Bengio, 2010) is used for parameter initialization. For the context-aware word classifier, the bidirectional LSTM has 512 hidden units. For AC-GAN, the CNN discriminator uses convolutional filters of size $\{3, 4, 5\}$ and 128 filters for each size. The LSTM generator and discriminator both have 512 hidden units. The generator is further pretrained on 82K privacy policy sentences using a 10K vocabulary. The coefficient λ_C is set to $\{-1, 0, 1, 2\}$ respectively for 'clear,' 'somewhat clear,' 'vague,' and 'extremely vague' categories. $\mathbf{v}$ is initialized as a binary vector, where an entry is set to 1 if it is one of the 40 cue words for vagueness (Bhatia et al., 2016). Word embeddings are initialized to their word2vec embeddings and are made trainable during the entire training process.

S1	... while we use **[reasonable]**$_{tp}$ **[efforts]**$_{fn}$ to protect your PII, we can not guarantee its absolute security.
S2	We use **[third-party]**$_{fn}$ advertising companies to serve **[some]**$_{tp}$ of the ads when you visit our web site.
S3	The **[information]**$_{fn}$ we obtain from **[those services]**$_{fn}$ **[often depends]**$_{fp}$ on your settings or their privacy policies, so be sure to check what those are.
S4	In the event of an insolvency, bankruptcy or receivership, **[personal data may]**$_{tp}$ also be transferred as a business asset.

Table 5: Examples of detected vague words in privacy policies. $[\cdot]_{tp}$ denotes true positive, $[\cdot]_{fp}$ is false positive, $[\cdot]_{fn}$ is false negative. All unmarked words are true negatives.

False Alarms POS Tag	Perc. (%)	Misses POS Tag	Perc. (%)
Adjective	**37.19**	Noun	**47.64**
Noun	35.24	Adjective	25.07
Verb	20.53	Verb	13.31
Adverb	4.63	Adverb	5.62
Determiner	1.72	Determiner	2.79

Table 6: The most frequent part-of-speech (POS) tags appeared in false alarms and misses of detected vague words.

6.2 Predicting Vague Words

We compare context-aware with context-agnostic classifiers on detecting vague words in privacy policy text. The goal is to test an important hypothesis: that vagueness is an intrinsic property of words, thus a word being vague has little to do with its context. Results are presented in Table 4.

Interestingly, context-agnostic classifier yields a high recall score (78.15%) despite it ignoring context. This result indicates word vagueness can be encoded in distributed word embeddings. However, the low precision (11.30%) suggests that context is important for fine-grained analysis. While it is possible for experts to create a comprehensive list of vague terms for assessing privacy policies, extra effort is required to verify the tagged vague terms. Using a context-aware classifier produces more balanced results, improving the F-score from 19.71% to 60.08%. Our findings suggest that the context information is necessary for detecting vague words.

In Table 5, we present examples of detected vague words. The nouns have caught our attention. The classifier misses several of these, including "efforts," "information," "services," per-

System	Sentence-Level P (%)	R (%)	F (%)
Baseline (Majority)	25.77	50.77	34.19
LSTM	47.79	50.06	47.88
CNN	49.66	52.51	50.18
AC-GAN (Full Model)	51.00	53.50	50.42
AC-GAN (Vagueness Only)	**52.90**	**54.64**	**52.34**

Table 7: Results on classifying vague sentences.

haps because there is no clear definition for these terminologies. In Table 6, we found nouns consist of 47.64% of all the miss-detected vague words, while adjectives consist of 37.19% of the false alarms. There is also an interesting phenomenon. In S3, "Information" and "those services" are considered more vague by humans than "often depends." However, if those terms are removed from the sentence, yielding "The [..] we obtain from [..] often depends on your settings or their privacy policies." In this case, the vagueness of "often depends" become more prominent and is captured by our system. It suggests that the degree of vagueness may be relative, depending on if other terms in the sentence are more vague.

6.3 Predicting Vague Sentences

In Table 7 we present results on classifying privacy policy sentences into four categories: clear, somewhat clear, vague, and extremely vague. We compare AC-GAN with three baselines: CNN and LSTM trained on human-annotated sentences, and a majority baseline that assigns the most frequent label to all test sentences. We observe that the AC-GAN models (using CNN discriminator) perform strongly, surpassing all baseline approaches. CNN shows strong performance, yielding an F-score of 50.92%. A similar effect has been demonstrated on other sentence classification tasks, where CNN outperforms LSTM and logistic regression classifiers (Kim, 2014; Zhang and Wallace, 2015). We report results of AC-GAN using the CNN discriminator. Comparing "Full Model" with "Vagueness Only," we found that allowing the AC-GAN to only discriminate sentences of different levels of vagueness, but not real/fake sentences, yields better results. We conjecture this is because training GAN models, especially with a multitask learning objective, can be unstable and more effort is required to balance the two objectives (L_S and L_C). Example sentences generated by AC-GAN are presented in Table 9.

% (Freq)	Clear	SomeC	Vague	ExtrV
Clear	**39.4** (477)	59.8 (723)	0.7 (8)	0.2 (2)
SomeC	12.4 (284)	**85.2** (1945)	2.4 (54)	0.0 (1)
Vague	3.4 (31)	89.6 (828)	**7.0** (65)	0.0 (0)
ExtrV	1.2 (1)	88.9 (72)	9.9 (8)	**0.0** (0)

Table 8: Confusion matrix for sentence classification. The decimal values are the percentage of system-identified sentences that were placed in the specified vagueness class. For example: the item in (row 1, col 2) conveys that 59.8% of sentences (absolute count is 723) identified by the system as "clear" were actually "somewhat clear" according to humans.

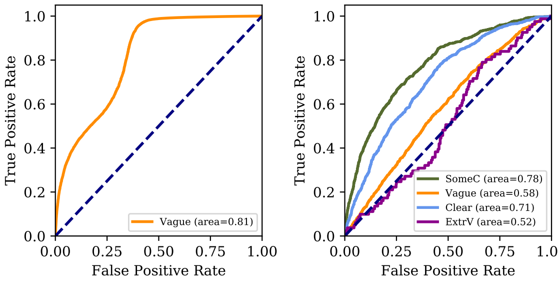

Figure 4: ROC curves for classifying vague words (left) and sentences (right).

Figure 4 shows the ROC curves of the four vagueness classes. Because the dataset is imbalanced, the ROC curves are more informative than F-scores. The "clear" and "somewhat clear" classes yield promising AUC scores of 0.71 and 0.78 respectively. The "vague" and "extremely vague" classes are more challenging. They are also the minority classes, consisting of 20.5% and 1.8% of the annotated data. Confusion matrix in Table 8 reveals that the majority of the sentences are tagged as "somewhat clear," while 7.0% of the vague sentences are tagged as vague. It suggests more annotated data may be helpful to enable the classifier to distinguish "vague" and "extremely vague" sentences. Interestingly, we found there is little correlation between the sentence vagueness score and sentence length (Pearson correlation r=0.18, p <0.001) while there is a relatively strong correlation (r=0.57, p <0.001) between sentence vagueness and the number of vague words in it. This finding verifies our hypothesis that vague words seem to increase the perceived sentence vagueness.

Lessons learned. We summarize some lessons learned from annotating and detecting vague content in privacy policies, useful for policy regulators, users and website operators. In general, privacy policies are suggested to:

- provide clear definitions for key concepts. Lacking definition is a major source of confu-

Clear	Our commitment to travian games uses paid services or send an order online.
	To learn how important anonymization it, we provide a separate medicare.
SomeC	Slate use certain cookies and offers.
	Visitors who apply us an credit card may sign up.
Vague	There may take certain incidents various offerings found on various topics; some or all individual has used.
	You may modify certain edit or otherwise delete certain features or a similar id will no longer.
ExtrV	Also, some apps may offer contests, sweepstakes, games or some community where necessary.
	If necessary, buying or clarify certain links, certain features of our site may place or some or some features may offer stack or unauthorized access some some functionality.

Table 9: Plausible sentences generated by AC-GAN. They exhibit different levels of vagueness. “SomeC” and “ExtrV” are shorthands for “somewhat clear” and “extremely vague.”

sion for the unfamiliar reader. Example concepts include personally identifiable information, personal (non-personal) information, third parties, service providers, subsidiaries, etc.

- suppress the use of vague words. There are on average 2.5 vague words per sentence in our corpus. The more vague words, the more likely the sentence is perceived as vague ($r = 0.57$);
- use sentences with simple syntactic structure to ease understanding. A sophisticated sentence with vague terms in it, e.g., “You may request deletion of your personal data by us, but please note that we may be required (by law *or otherwise*) to keep this information and not delete it...” appears especially confusing to readers.

7 Conclusion

In this paper we present the first empirical study on automatic detection of vague content in privacy policies. We create a sizable text corpus including human annotations of vague words and sentences. We further investigate the feasibility of predicting vague words and sentences using deep neural networks. Specifically we investigate context-agnostic and context-aware models for detecting vague words, and AC-GAN for detecting vague sentences. Our results suggest that a supervised paradigm for vagueness detection provides a promising avenue for identifying vague content and improving the usability of privacy policies.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions, which help to improve this paper. This work is in part supported by an ORC In-House Research Grant.

References

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A. Smith. 2012. Automatic categorization of privacy policies: A pilot study. Technical Report CMU-LTI-12-019, Carnegie Mellon University.

Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. 2016. A theory of vagueness and privacy risk perception. In *Proceedings of the IEEE International Conference on Requirements Engineering (RE)*.

Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2017. SaToS: Assessing and summarising terms of services from german webshops. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*.

Liwei Cai and William Yang Wang. 2018. KBGAN: Adversarial learning for knowledge graph embeddings. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4:357–370.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Ian Goodfellow. 2016. NIPS 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160* .

Jiatao Gu, Daniel Jiwoong Im, and Victor O.K. Li. 2018. Neural machine translation with gumbel-greedy decoding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Dominic Hyde. 2014. Sorites paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

Rosanna Keefe. 2000. *Theories of Vagueness*. Cambridge University Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fei Liu, Nicole Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *Proceedings of the AAAI Fall Symposium on Privacy and Language Technologies*.

Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A. Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*.

Aaron K. Massey, Jacob Eisenstein, Annie I. Anton, and Peter P. Swire. 2013. Automated text mining for requirements analysis of policy documents. *Proceedings of the 21st IEEE International Requirements Engineering Conference* .

Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *CoRR* abs/1411.1784.

Augustus Odena, Christopher Olah, and Jon Shlens. 2018. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the Thirty-fifth International Conference on Machine Learning (ICML)*.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. 2014. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia M. McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Law Technology Journal* 30(1).

Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, Cameron Russel, Florian Schaub, and Shomir Wilson. 2013. The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Technical Report CMU-ISR-13-119, Carnegie Mellon University.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stewart Shapiro. 2006. *Vagueness in Context*. Oxford: Oxford University Press.

Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, and Alexander M. Rush. 2016. Visual analysis of hidden state dynamics in recurrent neural networks. In *arXiv:1606.07461*.

Kees van Deemter. 2010. *Not Exactly: In Praise of Vagueness*. New York: Oxford University Press.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel R. Reidenberg, and Norman Sadeh. 2016a. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016b. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International World Wide Web Conference (WWW)*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*.

Ye Zhang and Byron C. Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *http://arxiv.org/abs/1510.03820v4* .