Preprint: In proceedings of IEEE International Conference on Big Data and Cloud Computing (BDCloud 2021)

The Effect of Text Ambiguity on creating Policy Knowledge Graphs

Anantaa Kotal
CSEE Department
University of Maryland, Baltimore County
Baltimore, USA
anantak1@umbc.edu

Anupam Joshi
CSEE Department
University of Maryland, Baltimore County
Baltimore, USA
joshi@umbc.edu

Karuna Pande Joshi
Information Systems Department
University of Maryland, Baltimore County
Baltimore, USA
karuna.joshi@umbc.edu

***Abstract*—A growing number of web and cloud-based products and services rely on data sharing between consumers, service providers, and their subsidiaries and third parties. There is a growing concern around the security and privacy of data in such large-scale shared architectures. Most organizations have a human-written privacy policy that discloses all the ways that data is shared, stored, and used. The organizational privacy policies must also be compliant with government and administrative regulations. This raises a major challenge for providers as they try to launch new services. Thus they are moving towards a system of automatic policy maintenance and regulatory compliance. This requires extracting policy from text documents and representing it in a semi-structured, machine-processable framework. The most popular method to this end is extracting policy information into a Knowledge Graph (KG). There exists a significant body of work that converts text descriptions of regulations into policies expressed in languages such as OWL and XACML and is grounded in the control-based schema by using NLP approaches. In this paper, we show that the NLP-based approaches to extract knowledge from written policy documents and representing them in enforceable Knowledge Graphs fail when the text policies are ambiguous. Ambiguity can arise from lack of clarity, misuse of syntax, and/or the use of complex language. We describe a system to extract features from a policy document that affect its ambiguity and classify the documents based on the level of ambiguity present. We validate this approach using human annotators. We show that a large number of documents in a popular privacy policy corpus (OPP-115) are ambiguous. This affects the ability to automatically monitor privacy policies. We show that for policies that are more ambiguous according to our proposed measure, NLP-based text segment classifiers are less accurate.**

***Index Terms*—privacy policy, ambiguity, knowledge graph, knowledge extraction, policy maintenance**

I. Introduction

Data sharing between consumers, service providers, and their subsidiaries and third parties are increasingly common these days [12]. Cloud-hosted services provide a low-maintenance alternative to hosting in-house technology. The cloud service architecture is built on a model of shared resources where there is a continuous flow of data. The resulting potential for inappropriate dissemination and usage of a given consumer's private information has raised concern among the public [3], prompting the creation of a plethora of data protection regulations like the Payment Card Industry Data Security Standard (PCI DSS) [11], the European Union's General Data Protection Regulation (GDPR) [1], and the Children's Online Privacy Protection Act (COPPA) [10]. A key element of this process is the extraction of these policies from text and representing them as Knowledge Graphs [18], [22], [23].

Companies must adhere to these regulations or risk expensive lawsuits or government sanctions [29], as recently seen in the case of the $5B FTC fine on Facebook. There is a significant effort to automate the process of policy maintenance and compliance [20], [21], [23]. Current data compliance policies exist as large text documents crafted by experts. Privacy policies represent a major class of compliance regulations. Privacy policies disclose all the ways that sensitive information such as personally identifiable information (PII) is collected, used, and stored, and how confidentiality is maintained. Privacy Policy documents help customers make informed decisions and provide transparency between users and organizations.

This raises a major challenge for a provider as they try to create new services by using or transforming data shared by the consumer. Once these services are deployed, it is nearly impossible to determine in real-time whether a policy violation has occurred, an issue that is becoming even more problematic as data moves at an extreme scale with high velocity within and across organizations. An ability to automatically monitor for the policy complaint use of shared confidential data on large public cyber-infrastructures like the cloud would simplify this process and protect companies and individuals from the repercussions of data protection regulation violations. Organizations are moving towards the automation of policy maintenance [6], [48], [50]. In addition to the organizational requirements, privacy policies must also comply with regulations. It is a tedious task to ensure that the privacy policies of an organization comply with all regulatory requirements. It is also prone to human error. Recently, efforts are being made to automate the process of regulation compliance in privacy policy [32], [23], [20] [14]. This reduces the administrative overhead. It also ensures that the privacy policies of an organization do not violate the regulations.

On the other side, once a customer shares the data with the organization, their ability to control access to it is limited. So a consumer needs to understand the security and privacy rules that govern the data they are sharing. The organizations disclose this information in privacy policies. Here again, policies that are presented as enforceable rules rather than pages of text are more useful. A Pew study in 2019 [4] showed that 81% of Americans don't feel they have control over data that is collected, and that the risk of collecting data outweighs benefits. 79% felt concerned about how the data collected was used, and 59% said they did not understand how it was being used. Only about a fifth of adults said that they always or often read these policies. Research conducted on 64 privacy policies by Jensen [19] determined the usability of online privacy policies and their accessibility to users. It was observed that the percentage of privacy policies that are accessible (considering the average reading grade of the common users) falls below 10%. This underlines the importance of the work done by Carminati [7], Mazzolini [34], [35], Kagal [25], [26] and others in expressing these policies as enforceable rules described in semantically rich languages.

A key step towards this goal is extracting meaningful information from a privacy policy and representing them as a Knowledge Graph (KG) in some semantically rich language such as XACML and OWL. New policy languages and models have been developed [9], [15], [37], [45] that enforce obligation requirements. There is also a significant effort to automate rule extraction and management from natural language text [2], [18], [20], [23], [38], [47]. This requires text segmentation and text classification, and the quality of text in a natural language document can affect the performance of these processes [40]. Kotal et al. [28] identified that privacy policies often have ambiguities and complex sentence constructions that make them harder to understand. Our approach uses linguistic properties beyond the ones described by those authors to measure the ambiguity of a policy. We also used these measures to show how the ambiguity of a privacy policy affects the creation of knowledge graphs or knowledge extraction from the policy documents.

Our key contribution is to show that the NLP-based approaches to extract knowledge from written policy documents and representing them in enforceable Knowledge Graphs fail when the text policies are ambiguous. In this paper, we describe a system to extract features from a policy document that affect its ambiguity and classify the documents based on the level of ambiguity present. We validate this approach using human annotators. We show that a large number of documents in a popular privacy policy corpus (OPP-115) are ambiguous. This affects the ability to automatically monitor privacy policies. We show that for policies that are more ambiguous according to our proposed measure, NLP-based text segment classifiers are less accurate.

The rest of the paper is organized as follows: In Section II, we discuss the background and motivation for our study. In Section III, we describe a way to extract linguistic features from a policy document to determine its ambiguity. In Section IV, we use the features extracted from a policy document to classify them into an ambiguity category. In Section V, we use 3 learning models for text segment classification in the privacy policies. We show how ambiguity in the document affects the performance of these tasks. We conclude the paper in Section VI.

II. Background and Related Work

Maintaining integrity and authenticity in data sharing has continued to be a topic of deliberation between experts. Consumers are concerned about their control over data, while companies must adhere to several data regulations to avoid potential lawsuits. The complex task of privacy policy management and the regulation task is typically done manually. However, human intervention is prone to error. There is a significant body of work that tries to find a more convenient solution to these tasks.

In response to the growing concern for individual privacy, several regulations like PCI-DSS [11], GDPR [1] and CALOPPA [10] have been introduced that define how an organization is allowed to control information. An organization might have to comply with multiple policy regulations based on its geographical and administrative location. It is often a complex task to ensure that the privacy policies of an organization comply with all regulatory requirements. Previous approaches automate the process of regulatory compliance in the privacy policy for cyber insurance services and Cloud data [20], [23]. This reduces the administrative overhead. It also ensures that the privacy policies of an organization do not violate the regulations.

A key step towards this goal is extracting meaningful information from a textual policy and representing it as a Knowledge Graph (KG) in some semantically rich language. This then allows reasoning to be done over policy rules. Carminati et al. [7] proposed an approach using RDF for policy specification and enforcement. Kagal et al. [25], [26] proposed an approach based on deontic logic and speech acts for policy description and maintenance and used RDF-S for the policy description language. Mazzoleni et al. [34], [35] and Takabi et al. [46] also proposed a semantic-based solution to policy language maintenance that uses XACML and OWL respectively. While policy languages and models were developed [9], [37] that enforce obligation requirements, we also need to consider the effectiveness of NLP approaches that target policy documents. There is also a significant effort to automate rule extraction and management from natural language text [2], [18], [47]. However, as Ravichander et al. [40] point out, there are challenges to automating knowledge extraction from policy documents. They argue that extracting rules requires text segmentation and text classification.

Jusoh [24] conducted a study to identify the most prominent challenges in NLP applications and summarised that "the complexity of the natural language itself, which is the ambiguity problems that occur in various levels of the language" is one of the key challenges to any NLP application. Ambiguity in a policy document makes it harder for a consumer to understand their rights. As we show, it also limits the effectiveness of creating knowledge graphs and thus, policy management on these documents. Previous works [28], [42], identify the

Imprecise Words	
Modal Words	may, might, can, could, would, likely
Condition Words	depending, necessary, appropriate, inappropriate, as needed, as applicable, otherwise reasonably, sometimes, from time to time
Generalization Words	generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things
Numeric Words	anyone, certain, everyone, numerous, some, most, few, much, many, various, including but not limited to
Probable Words	probably, possibly, optionally
Usable Words	adaptable, extensible, easy, familiar

TABLE I: Taxonomy of Imprecise Words

ambiguity problem in policy documents like privacy policies and Terms of Service (ToS).

There is a significant body of work that identifies lexical indicators that contribute to ambiguity in natural language (like English) [5], [16], [44]. Previous studies also identify and alleviate ambiguity in other natural language documents that are lexically similar to policy documents. Genova et al. [17] identifies the most frequently used indicators to detect ambiguity in software requirements text; Popescu et al. [39] propose an object-oriented approach to identifying such properties from a natural language text. Massey et al. [33] identifies ambiguity in regulatory requirements.

III. Measurable Properties of Policies

A key contribution of our work is showing that the ambiguity of policy documents affects the performance of knowledge extraction from privacy policy text. The first step to this is objectively determining the ambiguity in a policy. There are different approaches to identifying ambiguity in Natural Language [5], [16], [44]. Previous works on other technical documents define different linguistic properties of a document that affect its ambiguity [17], [28], [42]. In this section, we define 8 measurable features of a policy document that contribute to its ambiguity. We describe these features and explain how they can be extracted for a policy document.

Connective Words	
Copulative Words	and, both, as well as, not only, but also
Control Flow Words	if, then, while
Anaphorical Words	it, this, those

TABLE II: Taxonomy of Connective Words

1) **Frequency of Imprecise Words:** In the English Language, certain words are inherently inexact. Frequent use of such words can make documents **ambiguous**.
   For example, generalizing terms like “typically” or “generally” can be hard to interpret. Consider the following sentence taken from the Policy of a Web service provider, “The Services will be provided in a professional, timely and workman like manner by persons with the proper skill, training, and background, and consistent with generally accepted industry standards”[1]. The use of the word “generally” here makes it hard to understand the exact meaning. The average user of these Services, with no prior understanding of the law, can not understand what this policy entails.
   We define **Imprecise Words** as the words that are inexact in meaning. A measure for the quality of a policy document can be the frequency of imprecise words in the document.
   Previous studies attempt to identify such vague and imprecise terms that affect the quality of a technical document or policy [43]. We collect the words identified in previous studies as “imprecise” and create our taxonomy of “imprecise words”. This has been indicated in Table I.
   In our framework, we tokenized the words in a document and counted the number of “imprecise words” in the document. We got the frequency of imprecise words by dividing the count of imprecise words by the total count of words in the document.

2) **Frequency of Connective Words:** Connective words are used to link clauses or sentences in the English Language. They are important for the construction of meaningful sentences. However, overusing connective words increases the **complexity** of a document.
   Consider the following sentence taken from a Policy: “Like most online service providers, we collect information that web browsers, mobile devices, and servers typically make available, including the browser type, IP address, unique device identifiers, language preference, referring site, the date and time of access, operating system, and mobile network information.”[2]
   The word “and” has been used 3 times in the sentence to join multiple clauses. This is a difficult sentence to

[1] https://bit.ly/33FAckK
[2] https://bit.ly/2Gn32Ob

read. Related studies on textual requirements [17], [27] proposes measuring the frequency of connective words to evaluate the quality of text in Software Requirements. In our framework, we have created a taxonomy of connective words in Table II. We measured the frequency of these connective words in Policies. We used this measure in our estimate of the overall ambiguity of a document.

3) **Frequency of Polysemous Words:** Policy documents should have **clarity**. This means that the documents should be clear in their meaning, with no room for dubious interpretation.
Polysemous words are words that have multiple interchangeable meanings. The use of polysemous words, without further clarification or context, can lead to overall ambiguity of meaning. Hence, it is important to limit the frequency of polysemous words in a policy document.
Consider this example from the Policy of a cloud provider, "This Agreement commences on the last date of execution of the Order Form."[3] Here the word "execution" is a polysemous word and can mean "fulfilling an obligation" or "signing of the document". Without further context, we can't be sure of its intended meaning. However, we didn't consider polysemous words that have multiple meanings in different parts of speech (POS), e.g. the word "direct" has the following meanings: without intervening factors or intermediaries (adj.), with no one or nothing in between (adv.), aim (something) in a particular direction (verb), etc.
Consider the sentence: "You can clarify this kind of situation in your Privacy Policy by stating that while the tool you're developing doesn't collect any other personal information directly."[4] In this statement the POS category of "directly" is an adverb. Even though "direct" has multiple meanings, it is clear from the context what was meant here. Hence, we only consider words that have multiple meanings for the same parts of speech (PoS). We define such words, in the context of a document, as "polysemous".
The lexical database, Wordnet [36], is a commonly used tool for text analysis in English. In Wordnet, words are grouped into sets of cognitive synonyms (synsets). Each synset expresses a distinct concept. A word associated with more than one synset has multiple meanings.
In our framework, we tokenize the words in a policy document. We get the POS tag and synset association for each word. We prune the synset associations of a word by its POS tag. If a word has more than one such association, we flag it as 'polysemous'.
Let the count of polysemous Words in a document be P. Let the total count of words in that document be C. We measured the Frequency of Ambiguous Words as P/C This is one of our measures in estimating the overall ambiguity of a policy document.

4) **Readability Score:** Policies should have **high readability**. There are many existing Readability Tests, developed by Linguists. Readability tests are formulae that can be used to evaluate the **readability** of a text.
Most readability tests are built on the idea that longer sentences and words are more difficult to read. They use the count of syllables, words, and sentences in a text to give a measure of overall readability. Some words like "caterpillar" have multiple syllables, but they are commonly known. Hence, readers typically do not find any difficulty with these words. Hence, some readability tests overcome the effect of such words by referring to a list of words graded for difficulty.
The Dale–Chall readability formula [8] is a readability test that measures the comprehension difficulty that readers face when reading a text. It uses a list of 3000 words that groups of fourth-grade American students could reliably understand, considering any word not on that list to be difficult.
Based on the frequency of difficult words in a text, Dale-Chall readability formula assigns a readability score. The general formula is as follows:

$$0.1579\left(\frac{\text{difficult words}}{\text{words}}\times 100\right)+0.0496\left(\frac{\text{words}}{\text{sentences}}\right) \tag{1}$$

In our framework, we took the Readability Score of the document assigned by the Dale-Chall readability formula. This is one of the measures we used in estimating ambiguity in a policy document.

5) **Frequency of Punctuation:** Punctuation is used in a sentence to separate clauses and thoughts. This can help in clarifying the meaning. However, similar to connective terms, excess use of punctuation increases the **complexity** and reduces the **readability** of a document.
Consider the same example from before: "Such changes, revisions or modifications ("Changes") shall be effective immediately upon notice to you, which may be given by any means including, without limitation, posting on the Sites or by email."[5]
Here, a comma has been used multiple times to convey dissociated thoughts in the same sentence. This makes the sentence unnecessarily complex.
The frequency of punctuation in a document is easily measurable by parsing the text. In our framework, we use the frequency of punctuation as one of the indicators of ambiguity in the policy document.

6) **Frequency of Acronyms:** Another measurable property of a policy document is the frequency of Acronyms.

[3] https://bit.ly/33FAckK
[4] https://bit.ly/3krowZz
[5] https://bit.ly/2HXne9E

An acronym is an abbreviation of a long word or phrase using a subset of its characters. It is useful in shortening frequently referenced, verbose terms. It saves space and often is easier to remember. They are a useful linguistic device.

However, the overuse of acronyms can make a text complicated. Acronyms for Terms of Art, without proper definition, can be unknown to a reader. Consider this excerpt from a Policy, "any information defined as 'Special Categories of Data' under Article 9 of the GDPR (e.g. biometric or genetic data, information about one's religious beliefs, race, sex life or orientation."[6]

Here, GDPR is an acronym for "General Data Protection Regulation". It is often referenced in privacy policies. In the document quoted above, there is no further clarification as to what it means. An average reader might not know the context of such acronyms. Policy writers must take care to limit their use of such Acronyms.

In our framework, we use the frequency of Acronyms in a document as one of the measures of its overall ambiguity.

7) **Misspelled Words:** The use of correct spelling is essential to maintain the quality of any written text. Incorrect spelling makes it harder for readers to comprehend a document. It can also affect a customer's trust in the service provider.

   In our framework, we used the python SpellChecker library to search for misspelled words in a document. We tokenized the words in a document and removed all proper nouns before using the SpellChecker. For each document, we recorded the frequency of misspelled words for the total number of words in the document.

   We used the frequency of Misspelled words as one of our measures for ambiguity in a policy document.

8) **Correct Grammar:** Like correct spelling for words, correct grammar is essential to the integrity of a text. Our framework checked for the correct use of grammar in each sentence of a policy document. For this, we used the python Language-Check library. For each tokenized sentence, it matches to a parse tree to check for the correct syntax. For each document, we record the frequency of sentences with incorrect grammar for the total number of sentences in the document.

IV. Assessing Ambiguity in Policy Document

In Section III, we identified the measurable properties of a policy document related to its ambiguity. We used the techniques defined to extract the features of policy documents in the OPP-115 corpus. For each document in the corpus, we extracted the 8 features described above. The maximum and minimum values for each property in the OPP-115 corpus are reported in Table III. In the next section, we used the feature vector set for the documents in the OPP-115 corpus to rank the documents based on their ambiguity. In our previous study [28], we determined that it was ideal to have 3 categories of ambiguity in a policy document. Based on their ranking of ambiguity, we labeled them as:

1) Very Ambiguous
2) Somewhat Ambiguous
3) Not Ambiguous

Document Properties	**Max Value**	**Min Value**
Freq. of Imprecise Words	0.09	0.021
Freq. of Connective Words	0.076	0.025
Reading Complexity	10.2	6.21
Freq. of Polysemous Words	0.39	0.11
Frequency of Punctuation	0.15	0.05
Frequency of Acronyms	0.04	0
Misspelled Words	5.44	0.54
Correct Grammar	0.23	0.06

TABLE III: Max and Min Feature values extracted from OPP-115 corpus

The task is to determine the level of ambiguity in a policy using its extracted features. This would normally be a classification task. Using the features extracted in Section III, we classify the documents into one of the 3 categories. For a classification task, supervised learning is preferred. However, it requires a significant portion of the data to be labeled. This is not feasible for the larger datasets of policy documents. In this paper, we demonstrate a semi-supervised clustering algorithm that requires only partial labeling to classify the ambiguity in a policy text. In our clustering task, we use an algorithm that minimizes a linear combination of both cluster dispersion and cluster impurity measures. This allows us to extract and identify clusters in this data space that have less label mixing. To show that the semi-supervised clustering algorithm works better for this classification, we also include the results from a Support Vector Machine(SVM) for the same dataset.

In the following sections, we explain how the corpus was partially labeled and how it was then used for the classification of the complete dataset.

A. Dataset & Annotation

The OPP-115 Corpus (Online Privacy Policies, set of 115) [49] is a collection of 115 website privacy policies in natural language with annotations that specify data practices in the text. We used this well-established dataset for policy analysis in our experiments.

For the classification task, we needed at least partial labeling of the corpus. It is difficult to fully label documents in a large corpus, using volunteer annotators. We show here that even with a partially labeled dataset, we can get good results. To partially label the dataset, we used human subjects. We

[6] https://bit.ly/2JnLVg7

collected a group of 20 regular web service users. All users in this experiment had English education up to 10^{th} Grade or higher. According to CalOPPA, policies should be readable by someone with 8^{th} grade fluency or higher. Our pool of annotators consisted of graduate students in computer science and workers in the IT industry. Hence, they were a good representation of the average readers of web service policies and agreements.

We gave each annotator a set of 12 policy documents from our corpus. We asked our annotators to determine the level of ambiguity in a document and assign a score of [1-3] where 1 corresponds to a "Very ambiguous" document, 2 to a "Somewhat Ambiguous" document, and 3 to a "Not Ambiguous" document category. For each document, we polled the label assigned by each annotator. The documents were labeled with the category that had majority votes.

B. Semi-Supervised Clustering Algorithm

For a classification task, supervised learning is more reliable and makes use of available data. However, as discussed earlier supervised learning models require a significant portion of the dataset to be labeled, which is not a realistic assumption for our problem. We partially annotated the dataset as shown in IV-A. This is not enough for a supervised learning algorithm to work well for our dataset. To address this problem, we used a semi-supervised clustering technique.

Demiriz et al. [13] describe a semi-supervised algorithm for clustering that minimizes a linear combination of the cluster dispersion and cluster impurity measures. We use this algorithm in our work. The objective is to select $K > 2$ cluster centers, $m_k(k = 1, ..., K)$, that minimize the following objective function, where CD is Cluster Dispersion and CI is Cluster Impurity:

$$min_{m_k,(k=1,...,K)} \beta * CD + \alpha * CI \tag{2}$$

As in the K-means algorithm, each point is assumed to belong to the nearest cluster center as measured by Euclidean distance. Each non-empty cluster is assigned a class label corresponding to the majority class of points belonging to that cluster. The overall algorithm is as follows:

- Determine cluster centers
- Partition the labeled data by distance to closest cluster center.
- Find non-empty clusters, assign a label to non-empty clusters by majority class vote within them.
- Compute dispersion and impurity measures:
    1) Induction: Use labeled data.
    2) Transduction: Use labeled + unlabeled data.
- Prune clusters with few members.
- Reassign the points to final non-empty clusters.

The dispersion measure used is the mean square error (MSE), this quantity is defined as:

$$MSE = \frac{1}{N} \sum_{k=1}^{K} \sum_{x \epsilon C_k} ||x - m_k||^2 s \tag{3}$$

We used the Gini Index for the impurity measure of a cluster. The Gini Index of a certain cluster ($GiniP_j$) is computed as:

$$GiniP_j = 1.0 - \sum_{i=1}^{K} (\frac{P_{ij}}{N_j})^2 \tag{4}$$

The impurity measure of a particular partitioning into K clusters is:

$$impurity = \frac{\sum_{j=1}^{K} T_{P_j} * GiniP_j}{N} \tag{5}$$

C. Results & Validation

We ran the semi-supervised clustering algorithm described in IV-B to label the rest of the dataset and assign a qualitative category to each of the 115 policy documents in the corpus. The number of documents from the corpus in each category is given in V. To show that the semi-supervised clustering algorithm works better for this classification, we also used Support Vector Machine(SVM) to classify the documents in the same dataset. A comparison of the performance of the two learning models is provided below.

To validate that our model works, we used the group of 20 regular web service users (Section IV-A) for validation. We gave each validator a set of 10 policy documents from our corpus. This set of documents is mutually exclusive from the set used for annotation. We asked our validators to assign each document to assign a level of [1-3] where 1 corresponds to a "Very ambiguous" document, 2 to a "Somewhat Ambiguous" document, and 3 to a "Not Ambiguous" document category. For each document, we polled the category assigned by each validator. The documents were labeled with the category that had majority votes. We then compared the category assigned for each document by our validators with the category assigned by our model.

The results of this experiment are shown in Table IV. Additionally, to show that the semi-supervised clustering algorithm works better for this dataset than a supervised algorithm, we also include the results of a Support Vector Machine(SVM) for the same dataset. For 8 out of the 10 policies in our Validation set, the quality category assigned by our algorithm was the same as the majority category assigned by Human Validators. For 2 policies, the quality category assigned by our algorithm differed from the majority category assigned by Human Validators by a single category rank. The F1-score of both Semi-supervised clustering and SVM for each ambiguity class is provided in Table VI. For each class, the semi-supervised clustering algorithm is more accurate than the SVM. The average F1-score for the semi-supervised clustering algorithm is 0.8.

V. Effect of Text Ambiguity on Creating Policy Knowledge Graph

In previous sections, we showed that a significant number of privacy policies for web and cloud-based services, such as in OPP-115 [49], are ambiguous. This makes the document

Privacy Policy Website	Majority Category assigned by Human Validators	Document Classification by Semi-Supervised Clustering Algorithm	Document Classification by SVM
google.com	Not_Ambiguous	Not_Ambiguous	Very_Ambiguous
timeinc.com	Somewhat_Ambiguous	Somewhat_Ambiguous	Very_Ambiguous
zacks.com	Not_Ambiguous	Not_Ambiguous	Very_Ambiguous
msn.com	Not_Ambiguous	Somewhat_Ambiguous	Very_Ambiguous
voxmedia.com	Somewhat_Ambiguous	Somewhat_Ambiguous	Very_Ambiguous
taylorswift.com	Somewhat_Ambiguous	Somewhat_Ambiguous	Somewhat_Ambiguous
abita.com	Very_Ambiguous	Somewhat_Ambiguous	Somewhat_Ambiguous
steampowered.com	Somewhat_Ambiguous	Somewhat_Ambiguous	Somewhat_Ambiguous
sidearmsports.com	Very_Ambiguous	Very_Ambiguous	Very_Ambiguous
tangeroutlet.com	Very_Ambiguous	Very_Ambiguous	Very_Ambiguous

TABLE IV: Validation of Document Classification by Semi-Supervised Clustering Algorithm and SVM against Majority Category assigned by Human Validators

Ambiguity Class	Number of Documents
Not_Ambiguous	47
Somewhat_Ambiguous	36
Very_Ambiguous	32

TABLE V: Distribution of Documents in each at each Ambiguity Class by Semi-Supervised Clustering Algorithm

Ambiguity Class	Semi-Supervised Clustering	SVM
Not_Ambiguous	1	0
Somewhat_Ambiguous	0.75	0.57
Very_Ambiguous	0.67	0.31

TABLE VI: F1-score of Semi-Supervised Clustering vs SVM for each Ambiguity Class

difficult for customers to understand and harder for machines to process. In this section, we show how it affects knowledge extraction from privacy policies. Previous works [31], [40] describe that text segment classification is a critical part of automated information extraction from policy text for populating Knowledge Graphs. We show that the classification of text segments is significantly harder for more ambiguous policies. We run the same tasks like those described in [31], requiring the segmentation of policies into data practices of relevance to privacy policies, described next. We used the same machine learning approaches – LRs, SVMs, and CNNs. That said, we do not know what hyperparameters they used, so our models might be slightly different for CNNs. We analyze the performance of classification models in each ambiguity category of documents. We show that classification algorithms have less accurate results for policies with high ambiguity.

A. Dataset

The rules or policies in a policy document or schema can be grouped by the data events that they are addressing. For example, policies talk about the choices that users have in sharing their data as well as the security measures that the organizations use to protect that data. It is useful to distinguish between the kind of event that a segment in the policy is addressing as a prelude to extracting a rule and populating a knowledge graph from it. It is thus necessary to identify text segments in a privacy policy and classify the "data practice" category of the segment.

In this experiment we used the OPP-115 corpus [49] which consists of 115 privacy policies of popular websites. They have been annotated by domain experts. The policy annotation scheme was developed to capture the data practices specified by policies. The policies were divided into paragraph-length **segments** for annotators to read. For each segment, an annotator labeled zero or more data practices from each of the following categories:

1) **First Party Collection/Use**: how and why a service provider collects user information.
2) **Third Party Sharing/Collection**: how user information may be shared with or collected by third parties.
3) **User Choice/Control**: choices and control options available to users.
4) **Data Security**: how user information is protected.
5) **Policy Change**: if and how users will be informed about changes to the privacy policy.
6) **Do Not Track**: if and how Do Not Track signals for online tracking and advertising are honored.
7) **Other**: additional sub-labels not covered by the other categories.

Data Practice Categories	Textual Quality of Document								
	Very Ambiguous			Somewhat Ambiguous			Not Ambiguous		
	LR	**SVM**	**CNN**	**LR**	**SVM**	**CNN**	**LR**	**SVM**	**CNN**
First Party Collection/Use	0.70	0.71	0.68	0.71	0.76	0.77	0.75	0.77	0.77
Third Party Sharing/Collection	0.64	0.67	0.59	0.70	0.75	0.72	0.71	0.71	0.79
User Choice/Control	0.53	0.53	0.56	0.61	0.70	0.66	0.75	0.76	0.75
Data Security	0.56	0.65	0.68	0.63	0.68	0.77	0.67	0.76	0.77
Policy Change	0.48	0.69	0.56	0.71	0.62	0.68	0.61	0.76	0.78
Do Not Track	0.48	0.57	0.48	0.63	0.59	0.67	0.71	0.67	0.68
Other	0.76	0.76	0.74	0.75	0.79	0.79	0.77	0.83	0.81
Average	0.59	0.65	0.61	0.67	0.7	0.72	**0.71**	**0.75**	**0.76**

TABLE VII: F-1 Score for Data Practice Classification of Policy Documents

B. Experimental Setup

We now show that the ambiguity of the documents directly affects the performance of the classification task, irrespective of the learning model. We report the performance of the learning models in each of the three categories of the documents w.r.t. textual quality. We show that for documents labeled "Very Ambiguous" by our measure, the performance of the classification task is worse, irrespective of the underlying learning algorithm and the data practice category when compared to documents labeled "Not Ambiguous".

The dataset consisted of 3,792 segments from 115 privacy policies. We represented the text of each segment as a dense vector using Paragraph2Vec [30] and the GENSIM toolkit [41]. This approach exploited semantic similarities between words in the vocabulary of privacy policies, acknowledging that the vocabulary in this domain is specialized but not completely standardized. We used 3 learning models to machine annotate the documents in the OPP-115 corpus. They are as follows:

- **Logistic Regression (LR)**
- **Support Vector Machine (SVM)**
- **Convolutional Neural Network (CNN)**

C. Experimental Results

We analyzed the f1-score of each learning model for a data practice category in each quality category of the OPP-115 corpus. The results from our experiment are detailed in Table VII. For each data practice category, the performance of every learning model improved or remained the same as the quality of the documents improved. Looking at the average F1-score of each learning model, the score improved for all 3 learning models from the ambiguity category "Very ambiguous" to "Not ambiguous". The differences were particularly stark in categories like "Third Party Sharing", "Data Security" and "Do Not Track", likely to be of most interest to a consumer. This shows that ambiguity in a policy affects the performance of text classification tasks. The data practice classification is one of the key steps to populating knowledge graphs from policy documents. This indicates that extracting enforceable policy rules from ambiguous documents would be much harder than from documents that are not ambiguous. It also shows that our proposed measure of ambiguity is useful in capturing this distinction, and so can be used by policy creators to see upfront if their proposed policy text should be revised.

VI. Conclusion

Cloud and web-based services rely on data sharing between consumers, service providers, and their subsidiaries and third parties. In fact, the sharing and transforming of data are central to their services and revenue generation plans. However, there is a concern for security and privacy in large-scale resource and data sharing models. These organizations have privacy policies that disclose all the ways that data is collected, shared, and used. Manually maintaining the policy documents and ensuring their compliance with policy regulations can be tedious and prone to error. Organizations are invested in automating the process of policy maintenance. As demonstrated by previous works, this requires populating Knowledge Graphs from privacy policies. Ambiguity in natural language text has been identified as a challenge to machine-assisted text analysis or NLP. In this paper, we show that ambiguity in privacy policies affects knowledge extraction from these documents. For this, we defined an objective set of properties for measuring ambiguity in the privacy policy and classify the policy documents from the measured properties. We validated our method with human annotators. We then showed that knowledge extraction is harder from more ambiguous policies. This study helps us understand the linguistic properties of human-written policy documents and helps in the ongoing effort to automate policy maintenance and regulatory compliance for large-scale online services. Our approach will help policy writers create a higher quality of policies that are more transparent to customers and more effective for automated management.

VII. Acknowledgement

This research was partially supported by a DoD supplement to the NSF award 1747724, Phase I IUCRC UMBC: Center for Accelerated Real time Analytics (CARTA). We would like to thank all the survey respondents for their participation and contribution to our study.

References

[1] General data protection regulation (gdpr) – official legal text. https://gdpr-info.eu/, 2013. (Accessed on 02/24/2021).

[2] Manar Alohaly, Hassan Takabi, and Eduardo Blanco. Towards an automated extraction of abac constraints from natural language policies. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 105–119. Springer, 2019.

[3] Clark D Asay. Consumer information privacy and the problems (s) of third-party disclosures. *Nw. J. Tech. & Intell. Prop.*, 11:xxxi, 2012.

[4] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information — pew research center. https://pewrsr.ch/3ywnkLl, November 2019. (Accessed on 02/24/2021).

[5] Daniel M Berry. Ambiguity in natural language requirements documents. In *Monterey Workshop*, pages 1–7. Springer, 2007.

[6] Michael A Bush and Robert Brandt. Automation system access control system and method, February 3 2015. US Patent 8,949,970.

[7] B. Carminati, E. Ferrari, and B. Thuraisingham. Using rdf for policy specification and enforcement. In *Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004.*, pages 163–167, 2004.

[8] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[9] Omar Chowdhury, Andreas Gampe, Jianwei Niu, Jeffery von Ronne, Jared Bennatt, Anupam Datta, Limin Jia, and William H Winsborough. Privacy promises that can be kept: a policy analysis method with application to the hipaa privacy rule. In *Proceedings of the 18th ACM symposium on Access control models and technologies*, pages 3–14, 2013.

[10] Federal Trade Commission et al. Children's online privacy protection rule ("coppa"). *Retrieved on September*, 16, 2016.

[11] PCI Security Standards Council. Official pci security standards council site - verify pci compliance, download data security and credit card security standards. https://bit.ly/3ywnthR, 2016. (Accessed on 02/24/2021).

[12] David Deming. Balancing privacy with data sharing for the public good - the new york times. https://www.nytimes.com/2021/02/19/business/privacy-open-data-public.html, February 2021. (Accessed on 02/25/2021).

[13] Ayhan Demiriz, Kristin P Bennett, and Mark J Embrechts. Semi-supervised clustering using genetic algorithms. *Artificial neural networks in engineering (ANNIE-99)*, pages 809–814, 1999.

[14] Lavanya Elluri, Karuna Pande Joshi, and Anantaa Kotal. Measuring semantic similarity across eu gdpr regulation and cloud privacy policies. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3963–3978. IEEE, 2020.

[15] Tim Finin, Anupam Joshi, Lalana Kagal, Jianwei Niu, Ravi Sandhu, William Winsborough, and Bhavani Thuraisingham. R owl bac: representing role based access control in owl. In *Proceedings of the 13th ACM symposium on Access control models and technologies*, pages 73–82, 2008.

[16] Alexander Franz. *Automatic ambiguity resolution in natural language processing: an empirical approach*, volume 1171. Springer Science & Business Media, 1996.

[17] Gonzalo Génova, José M Fuentes, Juan Llorens, Omar Hurtado, and Valentín Moreno. A framework to measure and improve the quality of textual requirements. *Requirements engineering*, 18(1):25–41, 2013.

[18] Michael Good and Benjamin Coffinberger. Method and system for translating natural language policy to logical access control policy, December 31 2020. US Patent App. 16/788,579.

[19] Carlos Jensen and Colin Potts. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478, 2004.

[20] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin. Semantic approach to automating management of big data privacy policies. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 482–491. IEEE, 2016.

[21] Karuna Pande Joshi and Agniva Banerjee. Automating privacy compliance using policy integrated blockchain. *Cryptography*, 3(1):7, 2019.

[22] Karuna Pande Joshi and Srishty Saha. A Semantically Rich Framework for Knowledge Representation of Code of Federal Regulations (CFR). *Digital Government: Research and Practice*, December 2020.

[23] Ketki Joshi, Karuna Pande Joshi, and Sudip Mittal. A semantic approach for automating knowledge in policies of cyber insurance services. In *2019 IEEE International Conference on Web Services (ICWS)*, pages 33–40. IEEE, 2019.

[24] Shaidah Jusoh. A study on nlp applications and ambiguity problems. *Journal of Theoretical & Applied Information Technology*, 96(6), 2018.

[25] Lalana Kagal, Tim Finin, and Anupam Joshi. A Policy Based Approach to Security for the Semantic Web. In *2nd International Semantic Web Conference (ISWC2003)*, September 2003.

[26] Lalana Kagal, Tim Finin, and Anupam Joshi. A Policy Language for A Pervasive Computing Environment. In *IEEE 4th International Workshop on Policies for Distributed Systems and Networks*. June 2003.

[27] Joseph E Kasser. The first requirements elucidator demonstration (fred) tool. *Systems engineering*, 7(3):243–256, 2004.

[28] Anantaa Kotal, Karuna Pande Joshi, and Anupam Joshi. Vicloud: Measuring vagueness in cloud service privacy policies and terms of services. *UMBC Information Systems Department*, 2020.

[29] Pravin Kothari. Multinationals face unique challenges for data privacy and security compliance - cpo magazine. https://www.cpomagazine.com/data-protection/multinationals-face-unique-challenges-for-data-privacy-and-security-compliance/, September 2018. (Accessed on 02/23/2021).

[30] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[31] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. Towards automatic classification of privacy policy text. *School of Computer Science Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-ISR-17-118R and CMULTI-17-010*, 2018.

[32] Abhishek Mahindrakar and Karuna Pande Joshi. Automating gdpr compliance using policy integrated blockchain. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 86–93. IEEE, 2020.

[33] Aaron K Massey, Richard L Rutledge, Annie I Antón, and Peter P Swire. Identifying and classifying ambiguity for regulatory requirements. In *2014 IEEE 22nd international requirements engineering conference (RE)*, pages 83–92. IEEE, 2014.

[34] P. Mazzoleni, E. Bertino, B. Crispo, and S. Sivasubramanian. Xacml policy integration algorithms: Not to be confused with xacml policy combination algorithms! In *Proceedings of the Eleventh ACM Symposium on Access Control Models and Technologies*, SACMAT '06, page 219–227, New York, NY, USA, 2006. Association for Computing Machinery.

[35] Pietro Mazzoleni, Bruno Crispo, Swaminathan Sivasubramanian, and Elisa Bertino. Xacml policy integration algorithms. 11(1), February 2008.

[36] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[37] Qun Ni, Elisa Bertino, and Jorge Lobo. An obligation model bridging access control policies and privacy policies. In *Proceedings of the 13th*

ACM symposium on Access control models and technologies, pages 133–142, 2008.

[38] Aritran Piplai, Priyanka Ranade, Anantaa Kotal, Sudip Mittal, Sandeep Nair Narayanan, and Anupam Joshi. Using knowledge graphs and reinforcement learning for malware analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2626–2633. IEEE, 2020.

[39] Daniel Popescu, Spencer Rugaber, Nenad Medvidovic, and Daniel M Berry. Reducing ambiguities in requirements specifications via automatically created object-oriented models. In *Monterey Workshop*, pages 103–124. Springer, 2007.

[40] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*, 2019.

[41] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Citeseer, 2010.

[42] Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S170 Table 2, 2016.

[43] Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190, 2016.

[44] Roger C Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631, 1972.

[45] Nitin Kumar Sharma and Anupam Joshi. Representing attribute based access control policies in owl. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 333–336. IEEE, 2016.

[46] Hassan Takabi and James BD Joshi. Semantic-based policy management for cloud computing environments. *International Journal of Cloud Computing*, 1(2-3):119–144, 2012.

[47] Ronald C Turner. Proposed model for natural language abac authoring. In *Proceedings of the 2nd ACM Workshop on Attribute-Based Access Control*, pages 61–72, 2017.

[48] Matthias Vogel, Bernhard Drittler, and Markus Kupke. Managing access control information, March 25 2008. US Patent 7,350,237.

[49] 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, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, 2016.

[50] Zhiyi Zhang, Yingdi Yu, Sanjeev Kaushik Ramani, Alex Afanasyev, and Lixia Zhang. Nac: Automating access control via named data. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pages 626–633. IEEE, 2018.