

FACULTY OF ENGINEERING AND INFORMATICS

DEPARTMENT OF APPLIED COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

COURSEWORK 2

AI'S POTENTIAL TO IMPROVE USER PRIVACY BY ANALYZING “TERMS AND CONDITIONS”.

OLAMIDE OLADIMEJI 23027790

## ABSTRACT

In this study, I use cutting-edge machine learning techniques to tackle the problem of interpreting complex terms and conditions (T&C) statements. I aim to create an artificial intelligence model that can automatically categorize T&C statements into 'important' and 'unimportant' categories, making it easier for users to understand. I use a dataset of different T&C statements and extract features using natural language processing methods like Word2Vec and TF-IDF. Because of its prowess in managing high-dimensional data, the Support Vector Machine (SVM) algorithm is used to train the model. My results show that the model can identify minute semantic variations in T&C statements, indicating a potential path toward improving accessibility and transparency in legal documents. This research has practical implications for consumer rights and legal compliance and is a contribution to automated document analysis.

TABLE OF CONTENTS

[ABSTRACT 1](#_Toc632196508)

[INTRODUCTION 4](#_Toc1450976673)

[PROBLEM DESCRIPTION 5](#_Toc611955998)

[PROJECT MOTIVATION 6](#_Toc862699171)

[OBJECTIVE 7](#_Toc1688589445)

[RESEARCH CHALLENGES 8](#_Toc608889204)

[REPORT STRUCTURE 8](#_Toc93948111)

[LITERATURE REVIEW 10](#_Toc1911211406)

[Related work 10](#_Toc718504501)

[Terms and Conditions 10](#_Toc243452924)

[Technical Review 10](#_Toc309479275)

[MODEL TRAINING AND BUILDING 12](#_Toc78513648)

[Classification algorithm 12](#_Toc1645159427)

[Important and Unimportant 12](#_Toc952134489)

[Important Statements 12](#_Toc1407312312)

[Unimportant Statements 13](#_Toc1283496395)

[Algorithms or Methods 14](#_Toc1663011378)

[Evaluation 16](#_Toc1526309185)

[DATA DESCRIPTION 17](#_Toc1656652266)

[Data Preparation 18](#_Toc515122079)

[MODEL PREDICTION AND PERFORMANCE 19](#_Toc1730036298)

[Critical analysis 20](#_Toc89896190)

[CONCLUSION 21](#_Toc1228229870)

[REFERENCES 22](#_Toc271322730)

[APPENDIX 23](#_Toc1363663115)

## INTRODUCTION

In the current digital age, where information is the new currency and digital interactions influence our day-to-day existence, protecting user privacy has become essential to using technology responsibly. Users are frequently asked to accept lengthy Terms and Conditions (T&C), which are essential in defining the relationship between service providers and users, with each click, download, or sign-up. But often, the opaque nature of these important documents, filled with long clauses and legalese that obscures their true implications on user rights and privacy is what makes them so difficult to read. The very agreements that govern the use of users' data are virtually inaccessible to the public, which presents a significant challenge for users trying to make informed decisions about their data. ￼

Artificial Intelligence (AI) provides a ray of hope in this environment because of its extraordinary speed at which it can process and analyze massive amounts of data. This report's main goal is to examine how AI can improve user privacy by revolutionizing the way that Terms and Conditions are analyzed. Through cutting-edge technologies, artificial intelligence (AI) can analyze and interpret legal texts' complexity and present it in a way that the average user can understand, highlighting important information.

## PROBLEM DESCRIPTION

Legal Complexity and User Comprehension

The principal obstacle tackled in this research is the intricate and frequently opaque character of Terms and Conditions linked to digital services and goods. The average user faces a significant comprehension barrier due to the legal jargon and lengthy prose used in these documents. Due to users' frequent agreement to terms without fully understanding their implications, particularly regarding privacy and data usage, this complexity results in a lack of informed consent. A social networking site's terms of service and privacy policy, which stated that it would share all your information with the NSA and that users would have to give their first-born child as payment, were accepted by 97% and 93% of 543 participants, respectively, according to an experimental study (Obar and Oeldorf-Hirsch, 2018). Another example was Zappos.com, a platform having over 24 million users (about the population of Texas), held by the court after numerous users suing them for leak in their information, it was discovered that Zappos.com reserved right to change the terms and condition at any time without informing the customers, because of their obscure nature, the court held that the agreement was unenforceable (Wikipedia, 2023).

The central problem is the users' inability to effectively process and understand the content of these agreements, but also there’s the inadequacy of current methods used by users to understand and navigate Terms and Conditions, it is time taking and quite stressful to the average user, according to a Deloitte survey of 2,000 American consumers, 91% of respondents agree to the terms and conditions of services without reading them. 97% of those in the younger age group of 18 to 34 agreed to the terms before reading, which is an even higher percentage (Cakebread, 2017).

## PROJECT MOTIVATION

It began in 2021, when my academic path was still anchored in electrical engineering, due to a curiosity and problem-solving mindset. Despite the differences in my field of study, I have always been drawn to the nexus between technology and useful answers to real-world problems. Around this time, the idea for this project began to take shape because of the realization that struck a chord with millions of digital users all over the world: the widespread disregard for reading and comprehending terms and conditions (T&C). The idea behind this project was inspired by a seemingly insignificant but important observation: users frequently ignore terms and conditions (T&Cs), which can have a negative impact on their rights and privacy. This observation raised a crucial query: Why not develop a program or model that would enable the average user to interpret and comprehend these complex legal documents? As I wrote down in my notes back then ‘a platform for going through the T&Cs, saving billions of people the stress, should be worth billions of dollars’. This thought, which I then jotted down in my personal journal, was more than a passing notion; it was a possible resolution to a common problem.

The idea remained dormant but alive in my collection of potential innovations, even though my initial searches in 2021 turned up no significant existing solutions to the problem. This idea resurfaced with renewed vigor when the coursework was given.

Therefore, the main goal of this project is to close the knowledge gap between laypeople and complex legal text by using AI to make the daunting task of sorting through T&C easy to understand and enjoyable. This progression from a simple concept jotted down in an engineering student's notebook to a fully developed computer science research project captures the spirit of innovation, the will to realize game-changing concepts regardless of one's educational or professional experience. This project is more than just an academic endeavor; it is an example of the unrelenting search for a solution that has the potential to completely change the way we engage with digital agreements, empowering users and improving transparency in the digital sphere.

## OBJECTIVE

The principal aim of this report is to showcase the proficient utilization of programming frameworks and cloud data analytics for the development of AI applications, with a particular emphasis on real-life problems.

To demonstrate how well cloud computing platforms can handle large datasets and intricate AI algorithms, we plan to explore each aspect of these platforms. The creation of a prototype AI system, a physical representation of the theoretical understanding applied to a dataset pertinent to our selected case study, is the focus of my research. This working prototype will demonstrate my comprehension of AI concepts while also offering a workable solution to an actual issue. In addition, I will focus on building and preserving an ML model pipeline with frameworks like TensorFlow or Scikit-Learn. This pipeline will be carefully constructed to make sure it adheres to the requirements for a function that is production-ready and embodies scalability, reliability, and efficiency. Analyzing and choosing the best AI methods and algorithms will be a crucial part of the research. Decisions will be both optimal and justified thanks to this process, which is based on the particular context and specifics of the issue at hand. Finally, I will discuss the research's ethical and legal implications. The discussion will not be vital as we negotiate the murky waters of user consent, data privacy, and the larger social ramifications of artificial intelligence in society.

## RESEARCH CHALLENGES

The inherent complexity and variability of the legal language used in T&C statements presented one of the main challenges encountered in this research. Much thought went into selecting and fine-tuning the Word2Vec and TF-IDF models to extract meaningful features from such high-dimensional text data. Furthermore, because "importance" in legal text is multifaceted, it was challenging to train an SVM model to accurately classify statements. It was important to make sure the model did not overfit when applied to previously unseen data. The task of interpreting the high-dimensional feature vectors in a way that yields actionable insights posed another difficulty. This task is essential for comprehending model decisions and for possible legal tech applications. Also, there were very limited research papers on this study.

## REPORT STRUCTURE

The project is divided into multiple essential stages: Data Preprocessing, which involves cleaning and standardizing the text data; Data Collection, where we gathered a wide variety of T&C statements; Model training involved training the SVM model on the processed data; feature engineering involved extracting meaningful features using Word2Vec and TF-IDF; and evaluation involved rigorously testing the model's performance against a set of metrics. Robustness and reproducibility were the main concerns in every step of the process.

## LITERATURE REVIEW

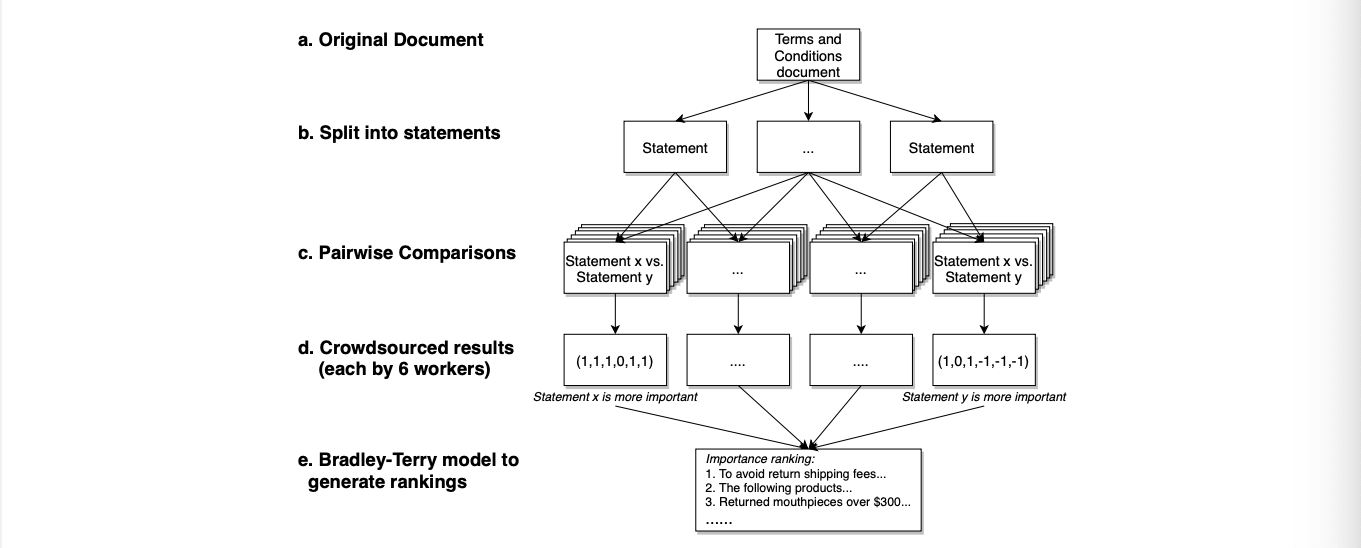
### Related work

In this section, I explained ‘terms and conditions’, previous methods used to make processing them easier, and clarify the approach used in collecting data from crowd workers.

### Terms and Conditions

For any product having ‘term and conditions’, we can define that to be the set of guidelines established by the policy holder / institution, so ideally, you must accept the ‘terms and conditions’ if you want access to the product, there is usually no other way around it.

### Technical Review

Fig 1 (Liu, Sun and Hong, 2021).

Workflow for crowdsourcing data

To determine which T&C document statements customers believe to be significant. First, the original T&Cs documents are divided into statements, or sentences, and then put together into pairs. Six different crowd workers are given each pair of statements to determine which is more important: 1 indicates that statement x is more important, 0 indicates that both statements are equally important, and -1 indicates that statement y is more important. The scores are then totaled, and the results are used to fit a Bradley-Terry model, which produces a ranking of the T&C statements' importance.

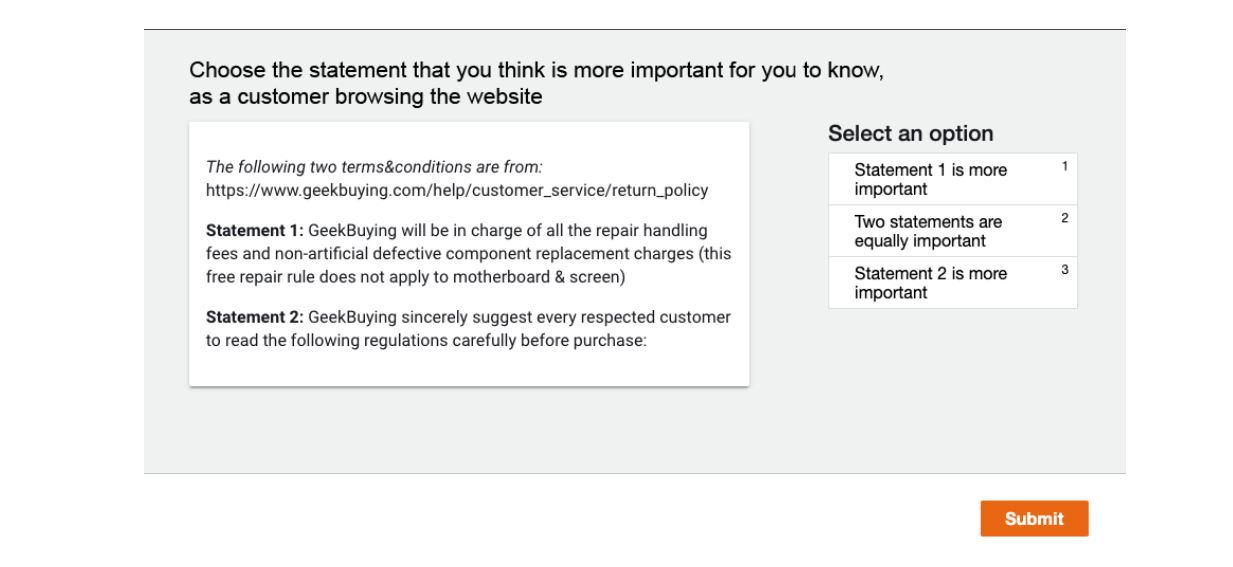


Fig 2 (Liu, Sun and Hong, 2021).

Crowd workers were presented with two randomly chosen statements from the same policy and asked to rank which of the two was more important. In our models, statements that were chosen more often than others would be regarded as having greater significance.

## MODEL TRAINING AND BUILDING

Prior to training the mode, the data was first divided into two classes: "important" and "unimportant" based on importance rankings produced by the crowdsourcing results. Since rankings and scores were produced comparative within the context of each document, they opted for the classification of important/unimportant statements over regression. Since a statement with a rank of 5 in Uniqlo's T&Cs is not always more important than a statement with a rank of 7 in Geekbuying's T&Cs, it was not possible to use statistical evidence to analyze the overall distribution and establish a threshold of important/unimportant. More specifically, each website's top T% of ranked statements was classified as “important,” while the remaining statements were classified as “unimportant”. The important threshold, T, of 10% was chosen for this model. Users are free to change it to suit their needs and improve the sensitivity, precision, and recall of their models.

Classification algorithm

Support Vector Machines (SVM) and pre-processed data were used to train a machine learning classification model. Support Vector Machine (SVM) was selected after a number of machine-learning algorithms, including Random Forest, Naive Bayes, and Logistic Regression, were tested. This was because SVM performed better overall. By utilizing the SVC API (Pedregosa et al., 2011), the SVM classifier was constructed and trained using Scikit-Learn's svm and the processed statement data, which were classified as "important" or "unimportant."

### Important and Unimportant

The classifications "important" and "unimportant" relate to the importance of statements found in documents that contain terms and conditions (T&Cs). The analysis and summarization of T&Cs, which are frequently lengthy and complex, depend heavily on this classification.

### Important Statements

These are the sections of the T&Cs that provide the user with important information. They could cover anything from user rights to privacy concerns to data usage policies, or any clauses that have the potential to substantially impact the user's comprehension of and decision to accept the terms. Finding such statements makes it easier to highlight significant points that users need to understand.

Unimportant Statements

On the other hand, these sections of the terms and conditions may include general, less important information that does not directly affect the rights or obligations of the user. This could include generic language, standard legalese, or broadly worded provisions that don't directly affect the user's experience or rights.

By grouping statements into these two categories, it becomes easier to separate the important information from the less important details, allowing users to make decisions without having to read the entire document. This classification is dynamic, meaning that the threshold (T) used to separate important from unimportant can be changed based on various factors, such as the needs of the user or the particular T&C context.

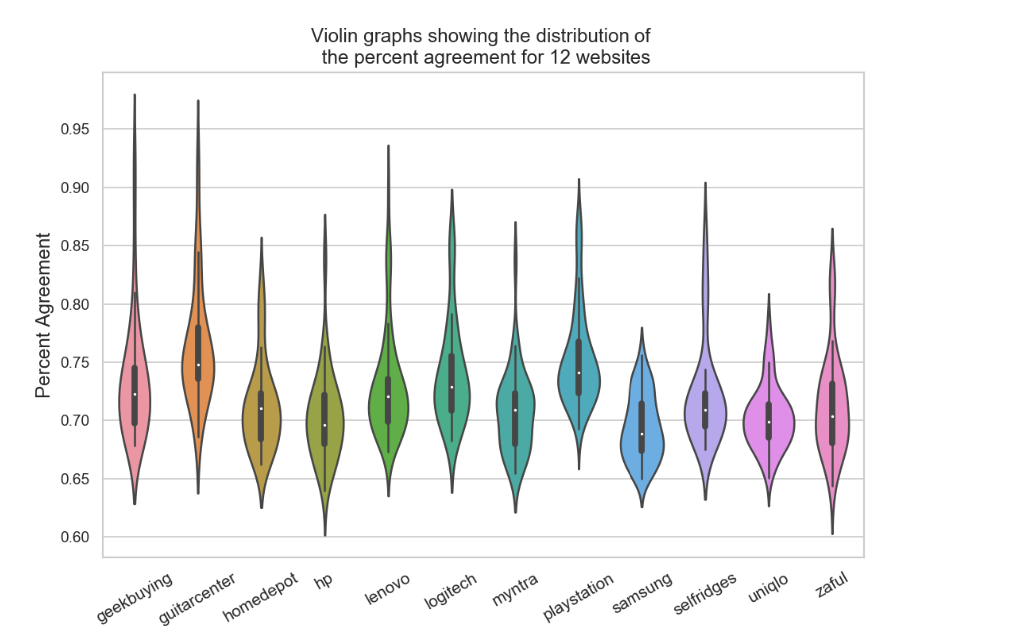


Fig 3 (Liu, Sun and Hong, 2021).

These are the percentage agreement graphs from 12 websites among 6 crowd workers for pairwise comparisons. For pairwise comparison tasks involving choosing which statement is more important, crowd workers came to a high level of agreement.

## Algorithms or Methods

Manual Labelers

Entails having legal professionals or other qualified people manually read and annotate privacy policies, marking problematic sections according to standards such as data sharing policies. Although this approach is very accurate and gives contextual understanding, it is laborious, less scalable, and may lead to subjective interpretations.

Natural Language Processing (NLP)

Employs sentiment analysis and keyword extraction techniques to find deceptive textual content and privacy-related practices. While natural language processing (NLP) is a useful tool for identifying patterns in vast amounts of text, its ability to comprehend legal jargon is limited and heavily dependent on the caliber of the training dataset.

The General Data Protection Regulation (GDPR) is a multi-chapter regulation consisting of 11 chapters and 173 recitals. Organizations primarily located in Europe are subject to GDPR. Nonetheless, the regulation might also be applicable to non-European organizations, for example, if these organizations provide goods or services to or keep an eye on individuals within Europe. A study by (Amaral CEJAS et al., 2021) focuses on applying AI to GDPR-compliant privacy policies to verify their completeness. To automate the process, it presents a conceptual model and a set of completeness criteria created by legal experts. 56 metadata types are included in the model, which is designed to capture the privacy-related provisions of GDPR. The method uses natural language processing (NLP) and machine learning to find GDPR-relevant data in privacy policies and compares it to completion standards. The approach showed excellent precision and recall when tested on 234 actual privacy policies, outperforming baselines based on keyword search-based methodologies by a significant margin.

A framework for analyzing privacy policies using knowledge graphs is called PoliGraph, and it is introduced in the research by (Chen et al., 2022). By using NLP to capture the relationships between data types, entities, and purposes, it can generate an organized representation of privacy policy texts. The paper provides new insights into the automated analysis of these intricate documents by outlining how PoliGraph can efficiently identify and summarize important privacy policy components. ￼

Deep Learning Techniques

Include using sequence modeling to comprehend dependencies in legal texts and training models such as CNNs or RNNs to classify clauses. These methods can be difficult to interpret, but they can be quite effective at learning intricate patterns and being flexible. They also demand a lot of computer power.

Using deep learning models for adaptability and learning complex patterns, natural language processing (NLP) for efficient preliminary analysis, and manual labelers for high-quality data annotation—a combined approach capitalizes on the advantages of each. By integrating AI-based techniques with manual analysis, this integrated method provides a more accurate and scalable way to identify concerning clauses in privacy policies.

"A Study on Attitudes and Perceptions Toward Privacy Policies" done by (Ibdah et al., 2021) investigates the opinions and attitudes of users regarding privacy policies. It uses data from 655 participants to identify factors that influence participants' willingness to read privacy policies and factors that act as roadblocks. The study assesses the influence on reading policies of prior experiences, including cyberattacks and online data sharing practices. The study evaluates users' comprehension of privacy policy content and the effect of technical jargon on readability. The results indicate that most users did not understand the content, despite a minority reporting difficulty in understanding.

### Evaluation

The analysis of "Terms and Conditions" documents using AI techniques has revealed both clear benefits and potential drawbacks. Despite the inherent complexities of legal text, the SVM model showed remarkable strength in handling high-dimensional data and produced dependable classifications. Notable benefits included its resistance to overfitting and its capacity to handle unbalanced datasets.

## DATA DESCRIPTION

Crowdsourced Data gotten by (Liu, Sun and Hong, 2021) contains these columns of data

Statement1: One sentence of T&C statement

Statement2: The T&C statement that was compared to statement1

StatementPosition1: position of the statement1 within the original T&C document

StatementPosition2: position of the statement2 within the original T&C document

TermsOfServiceId: ID of the selected T&C document

Webs: URL to the T&C document

Labels: a list of 6 crowd-workers' comparison results, 1 means statement1 was selected more important, -1 means statement2 was selected more important

Std: standard deviation of labels

Var: variance of labels

Maj\_percent: the percentage of the majority labels (e.g. [1,1,1,1, -1,-1] => 0.67)

vote: the aggregation of labels (e.g. [1,1,1,1,-1, -1] => 2)

param1: the raw Bradley-Terry ranking score of statement1 in this T&C document, the higher the score the more important it is ranked

param2: the raw Bradley-Terry ranking score of statement2 in this T&C document

rank1: the importance ranking of statement 1 in this T&C document, 1 means most important

rank2: the importance ranking of statement 2 in this T&C document

param\_diff: param1 - param2

rank\_diff: rank1 - rank2

website: the website where the T&C document is selected from

### Data Preparation

Data Association

To create a single text column, "statement1" and "statement2" were combined in the main association task. Consolidating the two statements was essential to guarantee that the analysis and model training took into account the context each statement provided.

Data Pre-processing:

Several steps were taken during the data cleaning process to standardize and streamline the text data:

Lowercasing: To preserve consistency and prevent duplication due to case differences, all text was converted to lowercase.

Punctuation Removal: To concentrate the analysis on the textual content, non-alphanumeric characters were eliminated.

After the text data had been cleaned, it was tokenized. This process turned each statement into a list of words, which is necessary for the next step of Word2Vec modeling.

Word2Vec Embeddings: Using the tokenized text data, we trained a Word2Vec model to discover word embeddings that capture semantic meanings according to the context of the dataset.

TF-IDF Vectors: Using the frequency and uniqueness of words across documents as a guide, we utilized the TF-IDF technique to highlight the significance of each term in the dataset.

Semantic context and term importance were blended to create a rich, nuanced feature set for every document by combining the Word2Vec and TF-IDF features.

Initial analysis

Key word distribution, recurring phrases, and the general organization of the terms and conditions statements were all examined in the preliminary analysis.

Data Characteristics or Features

For every statement, the combined Word2Vec and TF-IDF feature set represents a 100-dimensional vector that captures both term significance and semantic relationships. This high-dimensional feature space is useful for identifying minute variations in the terminology and conditions.

## MODEL PREDICTION AND PERFORMANCE

Classification was used to assess the predictive performance of the Support Vector Machine (SVM) model. This report includes a comprehensive examination of the precision, recall, and F1-score for every class in addition to the overall accuracy.

With a precision of 0.95 for the 'important' class, 95% of the instances that are predicted to be important are in fact important. The precision for the 'unimportant' class is 1.00, which indicates that there were no false positives in this category.

The model accurately identified 96% of all real important instances and 99% of all unimportant ones, demonstrating an equally high recall rate.

The model's balanced performance between precision and recall is highlighted by the F1-score, which is the harmonic mean of precision and recall. It is 0.95 for important and a perfect 1.00 for unimportant.

The support shows how many real instances of each class there are in the dataset (316 for important and 2953 for unimportant). This shows that there is a class imbalance, which the model has handled remarkably well.

The model's overall accuracy, which represents the percentage of all predictions that were accurate, is an astounding 99.11%.

The macro average of 0.97 for both precision and F1-score indicates that the model is performing exceptionally well across both classes, despite the imbalance. Even with the imbalance, the model is performing exceptionally well across both classes, as evidenced by the macro average of 0.97 for both precision and F1-score. This high degree of accuracy is also reflected in the weighted average, which accounts for the disparity in support among the classes.

These findings imply that the SVM model has a strong and highly reliable ability to distinguish between "important" and "unimportant" terms and conditions statements. Considering the intricate and subtle legal language found in T&C documents, such good performance suggests a well-trained model that could greatly improve the user's comprehension and navigation of these frequently tangled statements.

Such a model could be used in the application context to highlight significant clauses for end users, assisting them in making better decisions without having to sift through lengthy legal vocabulary. The user is reassured by the high precision and recall that the highlighted clauses are important and that very few important details are missed.

Critical analysis

An important development in legal text analytics is the creation of a cloud AI framework for terms and conditions interpretation. This method, which leverages a Support Vector Machine (SVM) model, demonstrates how machine learning can be used to automate and streamline the interpretation of intricate legal documents. The evaluation metrics show that the model performs well, which is indicative of a thorough understanding of the subtle qualities of the dataset attained through careful feature engineering and model training. But it's important to consider the constraints built into the training procedure, like possible biases in the crowdsourced data labeling process and the arbitrary nature of judging the 'importance' of legal statements.

The automated T&C statement classification affects how users interact with contracts, which they frequently sign without fully understanding. Although the model attempts to draw attention to important details, it is critical to make sure that the AI's interpretations are in line with legal knowledge. The use of AI for legal text interpretation must be carefully weighed against human oversight to avoid any misclassification that can cause users to miss important information and result in legal ramifications.

Concerns about data security and privacy are also raised by cloud AI, particularly when working with delicate legal documents. This approach demands safe handling and processing of data in the cloud environment, as well as rigorous compliance with data protection laws like GDPR.

## CONCLUSION

Artificial intelligence signals the beginning of a revolution in user empowerment and data privacy. The potential of artificial intelligence (AI) to analyze and clarify the sometimes complex "Terms and Conditions" documents that users frequently come across online has been investigated in this coursework. This research has shown that it is possible to classify complex legal language into easily interpreted categories by using sophisticated machine learning techniques, particularly Support Vector Machine (SVM) models. This improves transparency and user comprehension.

In conclusion, there is a great deal of room for AI to support user privacy within the framework of "Terms and Conditions". The progress made in this coursework is especially helpful in creating a setting where informed consent is the norm and user rights are respected, which adds to the larger conversation about AI's place in society. Building on this foundation, future work will enhance the model's functionality, broaden its application, and guarantee that the advantages of AI are achieved in a way that is morally and legally acceptable and, most importantly, user centered.

## REFERENCES

Obar, J.A. and Oeldorf-Hirsch, A. (2018). The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society, 23(1), pp.1–20. doi:https://doi.org/10.1080/1369118x.2018.1486870.

‌

Cakebread, C. (2017). You’re not alone, no one reads terms of service agreements. [online] Business Insider. Available at: https://www.businessinsider.com/deloitte-study-91-percent-agree-terms-of-service-without-reading-2017-11?r=US&IR=T.

Wikipedia. (2023). In re Zappos.com, Inc., Customer Data Security Breach Litigation [online] Available at: https://en.wikipedia.org/wiki/In\_re\_Zappos.com,\_Inc.,\_Customer\_Data\_Security\_Breach\_Litigation [Accessed 19 Dec. 2023].

‌

Amaral CEJAS, O., Abualhaija, S., Torre, D., Sabetzadeh, M. and Briand, L. (2021). AI-enabled Automation for Completeness Checking of Privacy Policies. IEEE Transactions on Software Engineering, pp.1–1. doi:https://doi.org/10.1109/tse.2021.3124332.

‌

Liu, X., Sun, A. and Hong, J.I., 2021. Identifying Terms and Conditions Important to Consumers using Crowdsourcing. arXiv preprint arXiv:2111.12182.

Ibdah, D., Lachtar, N., Raparthi, S.M. and Bacha, A. (2021). ‘Why Should I Read the Privacy Policy, I Just Need the Service’: A Study on Attitudes and Perceptions Toward Privacy Policies. IEEE Access, 9, pp.166465–166487. doi:https://doi.org/10.1109/access.2021.3130086.

‌

Chen, H., Rahmadi Trimananda, Athina Markopoulou and Jordan, S. (2022). PoliGraph: Automated Privacy Policy Analysis using Knowledge Graphs. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2210.06746.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

## APPENDIX

SOURCE CODE

COLAB FILE- [Link](https://colab.research.google.com/drive/1PbAxAT5nmShF29w6iBCYaILWtX6Pe8gj)

import pandas as pd

import re

import nltk

import spacy

from nltk.corpus import stopwords

# The spacy English model

!python -m spacy download en\_core\_web\_sm

nlp = spacy.load('en\_core\_web\_sm')

# Loading data

data = pd.read\_csv('clearterms\_data.csv')

# Converting to lowercase

data['statement1'] = data['statement1'].str.lower()

data['statement2'] = data['statement2'].str.lower()

# Removing punctuation

pattern = '[^\w\s]'

data['statement1'] = data['statement1'].apply(lambda x: re.sub(pattern, '', str(x)))

data['statement2'] = data['statement2'].apply(lambda x: re.sub(pattern, '', str(x)))

# Setting stopwords

nltk.download('stopwords')

stop\_words = set(stopwords.words('english'))

# Defining the lemmatization function

def lemmatize\_text(text):

doc = nlp(text)

return ' '.join([token.lemma\_ for token in doc])

# Applying lemmatization to the DataFrame columns

data['statement1'] = data['statement1'].apply(lemmatize\_text)

data['statement2'] = data['statement2'].apply(lemmatize\_text)

# Sorting the DataFrame based on the importance rank

data = data.sort\_values(by='importance\_rank', ascending=True)

# To determine the number of statements that constitute the top 10%

top\_10\_percent\_threshold = int(len(data) \* 0.10)

# Labelling the top 10% as 'important' and the rest as 'unimportant'

data['label'] = ['important' if idx < top\_10\_percent\_threshold else 'unimportant' for idx in range(len(data))]

from sklearn.feature\_extraction.text import TfidfVectorizer

# Concatenating texts columns

data['combined\_text'] = data['statement1'] + " " + data['statement2']

# Creating TfidfVectorizer object

vectorizer = TfidfVectorizer()

# To Fit and transform the combined text data

tfidf\_matrix = vectorizer.fit\_transform(data['combined\_text'])

from gensim.models import Word2Vec

import nltk

nltk.download('punkt')

from nltk.tokenize import word\_tokenize

# Concatenating both texts columns for each row and tokenizing

tokenized\_data = [word\_tokenize(text1 + " " + text2) for text1, text2 in zip(data['statement1'], data['statement2'])]

# Training a Word2Vec model using the tokenized data

word2vec\_model = Word2Vec(sentences=tokenized\_data, vector\_size=100, window=5, min\_count=1, workers=4)

#saved model

word2vec\_model.save("word2vec\_model.model")

from sklearn.feature\_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()

tfidf\_matrix = vectorizer.fit\_transform(data['combined\_text'])

from gensim.models import Word2Vec

import nltk

from nltk.tokenize import word\_tokenize

# Tokenizing and training Word2Vec

tokenized\_data = [word\_tokenize(text) for text in data['combined\_text']]

word2vec\_model = Word2Vec(sentences=tokenized\_data, vector\_size=100, window=5, min\_count=1, workers=4)

import numpy as np

# Initializing an array to store the combined vectors

combined\_vectors = np.zeros((len(tokenized\_data), 100)) # Assuming 100-dimensional Word2Vec vectors

for i, tokens in enumerate(tokenized\_data):

doc\_vector = np.zeros(100)

for token in tokens:

# To check if token is in both Word2Vec and TF-IDF vocabularies

if token in word2vec\_model.wv.key\_to\_index and token in vectorizer.vocabulary\_:

word\_vector = word2vec\_model.wv[token]

tfidf\_weight = tfidf\_matrix[i, vectorizer.vocabulary\_[token]]

doc\_vector += word\_vector \* tfidf\_weight

if np.linalg.norm(doc\_vector)!= 0:

doc\_vector /= np.linalg.norm(doc\_vector)

combined\_vectors[i] = doc\_vector

labels = data['label']

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(combined\_vectors, labels, test\_size=0.2, random\_state=42)

from sklearn import svm

model = svm.SVC()

model.fit(X\_train, y\_train)

from sklearn.metrics import classification\_report, accuracy\_score

y\_pred = model.predict(X\_test)

print(classification\_report(y\_test, y\_pred))

print("Accuracy:", accuracy\_score(y\_test, y\_pred))

import joblib

joblib.dump(model, 'svm\_model.pkl')