ANKARA YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ASSESSMENT OF GENERAL DATA PROTECTION REGULATION IN THE CONTEXT OF SMART CITY SOLUTIONS

M.Sc. Thesis by

Şeyma Nur TUNÇ

Department of Computer Engineering

December, 2021

ANKARA

ASSESSMENT OF GENERAL DATA PROTECTION REGULATION IN THE CONTEXT OF SMART CITY SOLUTIONS

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Şeyma Nur TUNÇ

December, 2021

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"ASSESSMENT OF GENERAL DATA PROTECTION REGULATION IN THE CONTEXT OF SMART CITY SOLUTIONS"** completed by **ŞEYMA NUR TUNÇ** under the supervision of **ASSIST. PROF. DR. M. ABDULLAH BÜLBÜL** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. M. Abdullah BÜLBÜL

Supervisor

Assist. Prof. Mücahid KUTLU

Jury Member

Assist. Prof. Fahreddin Şükrü TORUN

Jury Member

Prof. Dr. Sadettin ORHAN

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date:

Signature: ……
Name & Surname: ŞEYMA NUR TUNÇ

ACKNOWLEDGMENTS

Firstly, I would like to thank, Assist. Prof. Dr. M. Abdullah BÜLBÜL for his guidance, and support during my study. His recommendation and knowledge help me to motivate while writing my thesis.

I would like to thank, Assist. Prof. Dr. Rahime BELEN SAĞLAM for her guidance, and support during my study. Always she shared her knowledge fairly.

Finally, I can't admire the support of my husband, Oğuzhan TUNÇ. He never gave up his support during my study. I would not have been able to complete this thesis without him. This thesis is dedicated to my husband.

2021, 24 December **Şeyma Nur TUNÇ**

ASSESSMENT OF GENERAL DATA PROTECTION REGULATION IN THE CONTEXT OF SMART CITY SOLUTIONS

ABSTRACT

After the European Union's new General Data Protection Regulation (GDPR) became applicable in May 2018, concerns related to the legal compliance of several technologies including IoT devices, blockchain systems, or smart city applications have emerged considering the individual rights guaranteed by GDPR. This study aims to reveal the implications of GDPR on smart city applications. Smart city applications convert data coming from many interconnected devices to meaningful information and use them to raise the standards of daily life. To better understand how the smart city sector sees the challenges posed by GDPR, and how they address them, this study analyses privacy policies of smart city applications which are legal documents that explain what data is collected from users and how they are processed. These analyzes were made by taking into account nine main elements extracted from the GDPR, which may be important in the context of the smart city. These main topics are; Explicit Consent, Right to Withdraw Consent, Right to Erasure, Right to Data Portability, Right of Access, Storage Limitation, Transparency, Transfer of Data, Data Minimisation. According to these GDPR elements, personal data cannot be processed without the explicit consent of the user, users have the right to access the data at any time, to make copies of the data, to delete the data, and to withdraw their consent at any time. In this study, privacy policies of 68 applications are collected, web page-specific policies are eliminated, and 42 privacy policies are analyzed. Furthermore, this study contains results of the study aiming at automating the evaluation of GDPR compliance of privacy policies via information retrieval techniques. For this purpose, 70% of the privacy policies collected are used as a training set and 30% are used in the test set. During the training phase, the keywords, determined by running Tf-Idf over the selected privacy policies, are given as input to the system for the next test phase. The Okapi BM25 method calculated a score on sentences for each set of keywords, taking into account the given keywords. Considering the scores, the most appropriate sentences for each GDPR element were

automatically calculated. Then, the system was evaluated based on the results obtained. In addition, this method has been compared with fastText, which is one of the text classification methods, and the results are shared.

Keywords: Privacy, Privacy Policy, EU GDPR, Information Retrieval, Okapi BM25, fastText

AKILLI ŞEHİR ÇÖZÜMLERİ BAĞLAMINDA GENEL VERİ KORUMA YÖNETMELİĞİNİN DEĞERLENDİRİLMESİ

ÖZ

Avrupa Birliği'nin yeni Genel Veri Koruma Yönetmeliği'nin (GDPR) Mayıs 2018'de geçerli hale gelmesinden sonra, blok zincir sistemleri, IoT cihazları veya akıllı şehir uygulamaları da dahil olmak üzere birçok teknolojinin GDPR tarafından garanti edilen haklara yasal uyumu konusundaki endişeler ortaya çıktı. Bu çalışma, GDPR'ın akıllı şehir uygulamaları üzerindeki etkilerini ortaya çıkarmayı amaçlamaktadır. Akıllı şehir uygulamaları, birbirine bağlı birçok cihazdan gelen verileri günlük yaşam standartlarını yükseltmek için anlamlı bilgilere dönüştürüp kullanır. Bu çalışma kapsamında, akıllı şehir sektörünün GDPR' ın getirdiği zorlukları nasıl görüp ele aldığını anlamak için, akıllı şehir uygulamalarının kullanıcılarından hangi verileri toplandığını ve nasıl işlendiğini açıklayan gizlilik politikaları analiz edilmiştir. Bu analizler akıllı şehir bağlamında önemli olabilecek, GDPR içerisinden çıkarılan dokuz temel unsur göz önünde bulundurularak yapılmıştır. Bu unsurlar sırası ile; Kullanıcıdan Açık Onay Alma, Verilen Onaydan Vazgeçme Hakkı, Veri Silme Hakkı, Verileri Taşıma Hakkı, Verilere Erişim Hakkı, Verileri Saklama Sınırı, Açıklık, Veri Transferi, Minimum Veri İşleme olarak verilebilir. Bu temel prensiplere göre, kullanıcıların açık onayı olmadan verileri işlenemez ve kullanıcılar istedikleri zaman verilere erişim, verilerin kopyasını alma, silme ve istediği zaman verdiği onaydan vazgeçebilme hakkına sahiptir. Çalışma kapsamında, 68 uygulamanın gizlilik politikaları toplanmış, bunlardan bir kısmı sadece web sayfasına özgü olduğundan elenip, 42 tanesi detaylı incelenmiştir. Ayrıca bu çalışma bilgi geri getirme tekniğini kullanan, gizlilik politikalarının GDPR ile uyumluluğunu otomatik hale getirmeyi amaçlayan bir çalışmayı da içerir. Bu amaçla toplanan gizlilik politikalarının %70'i eğitim seti olarak, %30'u test setinde kullanıldı. Eğitim aşamasında, belirlenen gizililik politikaları üzerinden Tf-Idf çalıştırılarak belirlenen anahtar kelimeleri bir sonraki test aşamasında, sisteme girdi olarak verildi. Okapi BM25 yöntemi, verilen anahtar kelimeleri göz önünde bulundurarak, her bir anahtar kelime kümesi için, cümleler üzerinde bir puan hesaplandı. Puanlar göz önünde bulundurularak, her bir GDPR elementi için en uygun cümleler otomatik şekilde

hesaplanmış oldu. Daha sonra, elde edilen sonuçlar üzerinden sistem değerlendirildi. Ayrıca bu yöntem, metin sınıflandırma yöntemlerinden biri olan fastText ile karşılaştırılıp, sonuçlar paylaşılmıştır.

Anahtar Kelimeler: Gizlilik, Gizlilik Politikası, EU GDPR, Bilgi Getirimi, Okapi BM25, fastText

CONTENTS

NOMENCLATURE

Acronyms

GDPR	General Data Protection Regulation
EU	European Union
ICT	Information and Communication Technology
IoT	Internet of Things
RFID	Radio Frequency Identification
TP	True positive
TN	False negative
FP	False positive
FN	False negative

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 General View

The recent rapid developments in technology have caused many issues related to data protection. In May 2018, The European Union's General Data Protection Regulation (GDPR) [13] which became enforceable across the whole European Union (EU). It is a robust and new regulation concerning data protection. It aims to protect the privacy of any data subject in the EU as well as the data collected and processed there. While providing this protection, it is not taken into account where the data is. To achieve its set goals, the GDPR specifies rights of data subjects (here users of smart city applications), the lawful basis for processing personal data, and principles that should be considered by data controllers (the companies that run the smart city applications) while processing personal information.

Smart city applications transform the data, which is collected through IoT devices, into meaningful information to facilitate daily life. Since it also processes personal data, it is inevitable to comply with the GDPR. In this study, the compliance with the GDPR from the smart city perspective was researched, analyzed and the results were shared.

Firstly, assessments are done manually. Since manual evaluation is a very time-consuming task, there was a need to automate it. To automate this, an information retrieval technique has been applied and evaluation results of the system are discussed.

1.2 Scope of the Work

In this study, 42 privacy policies are collected to evaluate smart city applications' compliance with the GDPR. Firstly, GDPR elements that will be important within the scope of the smart city were determined. Then, sentences were extracted from

these collected documents according to the specified GDPR elements. The collected data were evaluated manually and the results are shared. Since this method takes a lot of time, a new method has been determined to make the evaluations automatically. Then, the success of this new method is compared and the results are presented.

1.3 Contribution of Thesis

This study investigates the GDPR compliance of smart city applications. For this purpose, the privacy policies accessed via smart city application's websites were evaluated in terms of key principles extracted from GDPR. In addition, this study aims to automatically analyze privacy policies by using the Okapi BM25 method. Most of the study as stated [18-21], generally well-known machine learning methods are used to train systems. On the other hand, in this study, a different approach is stated by using the scoring algorithm. The system aims to find relevant parts from the document to tag it while scoring extracted sentences. This information retrieval method was compared with another popular classification technique, fastText, and it was seen that it showed a more successful result. Also, this study contributes the data set for GDPR compliance of smart city context.

1.4 Thesis Organization

The rest of the thesis is organized as follows. The next section provides useful information about the background of this study. It gives information about what smart city, GDPR, the relationship between GDPR and smart city, and literature reviews. Another chapter gives information about what methodology is used in this study. Then, implementation details and results about manual assessment and automatization of the assessment are discussed. The thesis with conclusion comments is stated in the last chapter.

CHAPTER 2

BACKGROUND

2.1 Smart City

The smart city is one of the latest developments built on IoT technology, big data analysis, cyber-physical systems, and real-time control to make humans lives easier as stated in Figure 2.1. Monitoring the physical world in real-time, a smart city can provide an intelligent environment to both citizens and travelers in several contexts such as transportation, healthcare, or even entertainment. This monitoring is accomplished via IoT devices that are connected to transfer data over the network. This transferred data can be collected from several sources such as people, devices, sensors, RFID systems, cameras, microphones, or versatile wearable devices. Via heterogeneous networks, this information is transmitted to a control center which is a powerful computing system. This center process and analyze the collected data. Finally, this center makes the most appropriate decisions, then manipulates the urban operations with feedback components.

Several smart city applications have been implemented so far with the advancement of the IoT, cloud computing, and big data. Smart energy, that monitors energy generation, transmission, distribution, and consumption; smart environment, which provides to support a sustainable environment and comfortable climate; a small industry that optimizes industrial production and manufacturing; smart living, which manages various devices and services for comfortable home environment and finally smart service which enables public services and facilities to benefit people in many ways can be given as key applications.

In order to provide an insight about the variety of the information processed by those applications, it is worth covering some specific applications here. Smart Waste Management System is one of the smart city applications, the garbage cans have sensors to determine the fullness of the bin to make a decision [8]. Another example can be given as Smart Healthcare System which permits healthcare providers and

practitioners to gather, analyze patient information such as blood pressure, blood sugar, and sleeping patterns, etc. [9]. On the other hand, Smart Traffic Management Systems control the traffic to prevent congestion. This system is managed via monitoring registration data such as civil status (age, marital status), residence, or job as well as geo-coordinates which includes current GPS location [5].

As cities become smarter, applications raise a series of concerns and challenges in the context of security and privacy [10]. Since a huge amount of private information is collected and processed, those applications are open to privacy leakage and stealing information by attackers [11]. In this paper, we highlight those issues and explore how smart city applications provide solutions to guarantee data privacy through the lens of the GDPR.

Figure 2. 1 Smart City [12]

2.2 The GDPR: Overview and Related Elements

The General Data Protection Regulation (GDPR) was introduced to protect the personal data of its citizens living in the EU. The GDPR, which entered into force on May 25, 2018, made compliance mandatory in institutions that process personal data.

In this section, we provide definitions for a subset of GDPR elements that are important in the smart city context and listed in Figure 2.2.

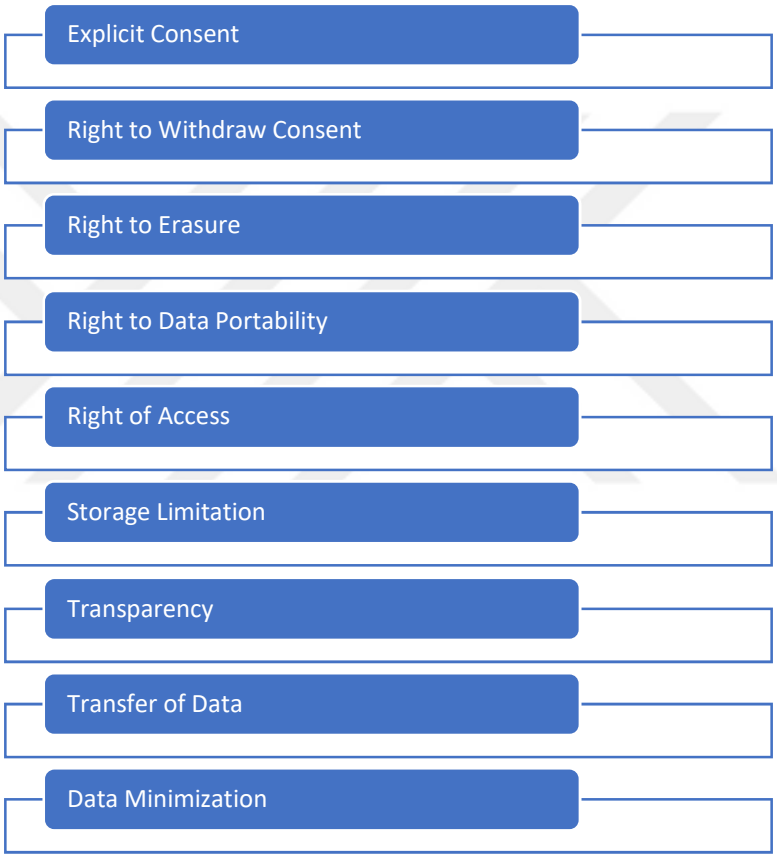

Figure 2. 2 GDPR elements of by the context of the smart city

2.2.1 Explicit Consent

The GDPR defines a set lawful basis to process any personal data, which includes acquiring explicit consent from the data subject. The consent should be freely given and it should be informed, specific, and a clear indication of the data subject's preferences regarding the processing of own personal data. The data controller must obtain explicit consent from the data subject and a data subject has the right to roll back consent at any time.

2.2.2 Right to Withdraw Consent

Article 7 in the GDPR clearly states that a data subject can withdraw content at any time. The regulation also makes it clear that giving and withdrawing consent should be equally easy.

2.2.3 Right to Erasure

Generally known as the right to be forgotten, the right to erasure requires data controllers to delete data when there is no longer a lawful basis for the processing or when the data subject withdraws consent and requests to delete personal information, or if the data is processed illegally.

2.2.4 Right to Data Portability

Individuals may request their data to be retrieved and reused across different services with this right. According to this rule, a data subject can receive a copy of personal data in the specified format and transfer it easily from one IT service to another in a secure and safe way, while preserving usability of it.

2.2.5 Right of Access

Article 15 allows Individuals to access their personal data processed by a system. Besides, they can get a copy of their personal data, GDPR also requires data controllers to provide individuals with the purposes of the processing.

2.2.6 Storage Limitation

Article 5.1 in the GDPR dictates that data controllers/processors must delete personal data when it is no longer necessary, and they should be clear about how long this personal data is kept in the system.

2.2.7 Transparency

According to Article 5, data controllers must process personal data in a transparent manner. They should be open and honest to people on how to handle their data.

2.2.8 Transfer of Data

GDPR clearly states that data controllers should give information to the data subject if they transfer data to third parties, third countries, or another place like outside the EU.

2.2.9 Data Minimization

The GDPR dictates that data controllers/processors have to give assurance that the processing of personal data is “sufficient, convenient, and minimum”

2.3 Smart City and GDPR

A smart city contains numerous physical devices connected to a common network – usually referred to as the Internet of Things (IoT) - to accumulate data and uses information gained from that data to manage assets, resources and services efficiently [1-2]. In other words, the data and technology used by smart cities enable them to make good decisions [3]. According to the results of recent studies, the investments in smart cities have increased day by day [4]. The development of big data and Information and Communication Technology (ICT) have a huge impact on this increase.

Despite their benefits in raising the standard of everyday life, smart city applications cause several concerns due to the huge amount of personal information they process and violation. Systems utilized in smart city applications such as video surveillance

raise concerns regarding the violation of confidentiality of personal information. Video surveillance systems make use of face recognition techniques to evaluate images captured by CCTV cameras [5]. The use of face biometrics, which is considered sensitive by GDPR, causes important privacy concerns [6]. In their study, Hasan et al. focused on bike-sharing programs which are widespread in smart city environments [7]. Researchers report that these systems collect visitor information with the identity of the user and make it public after anonymizing the data. It is argued that the dataset that is made publicly available will still be vulnerable to privacy violations even though it does not cover personally identifiable information. Researchers argue that it is possible to de-anonymize the data using the timing and location information of individuals [7].

The increase in the personal information collected by smart city applications has made it mandatory to comply with data protection regulations. The EU's GDPR is one of the most powerful regulations for data protection. The privacy and any collected or processed personal data of any data subject who locates in the EU are protected with this regulation.

2.4 Literature Review

The Literature Review section is organized under the two main titles, the first section states manual assessment-related studies, the second section includes studies on the automatization of these manual assessments.

2.4.1 The Perspective of Smart City Compatibility with GDPR

After the GDPR became mandatory for all organizations, smart city solutions were also affected by this regulation. Research on this subject also accelerated as a result of it.

In 2018, Goran Vojkovic indicated whether GDPR will have drawbacks against the acceleration of smart city solutions [14]. He stated not all smart city solutions violate GDPR, especially those do not include personal information in the process.

Additionally, he provides some advice for smart city solutions that need personal information. One of them is to provide pseudonymization.

Another study focuses to find a solution to many questions includes compatibility with GDPR [15]. They emphasize smart traffic management, real-time driving, and smart payment. To address the GDPR compatibility, they also provide a flowchart in the context of GDPR compliance.

Goo Yeon Lee et al. provide a way to collect personal information as a GDPR complaint in the IoT context [16]. Firstly, they specified the flow of personal information then they decide the main points that include this information. For these found points they share their implementation.

Michael M.Losavio, K.P. Chow, Andras Koltay and Joshua James in their study [17] assess the relationship between privacy and security in the IoT and smart city domain. They compare different nations by considering their digital forensic.

2.4.2 The Perspective of Automated Analyzing Privacy Policy

The manual evaluation of the privacy policies by considering GDPR compatibility is time-consuming because privacy policies are more comprehensive texts. Some researchers published related works to reduce this time.

Elisa Constante et.al. proposed some machine learning methods to automatize the privacy policies. In their work, they first decided to define categories that are used in the completeness step. These categories can be listed as; choice and access, data collection, data sharing, purpose, retention time, security, cookies, advertising, user location data, children, contact, policy change, trust seals, data storage location, links to external sites, safe harbor, California privacy rights [18]. They created a corpus with 64 privacy policies. After the data preprocessing step, they continue with machine learning step. For the machine learning step, they used Naïve Bayes, Linear Support Vector Machine, Ridge Regression, the k-Nearest Neighbor, Decision Tree, and Support Vector Machine. [18]

Welderufael B.Tesfay et. al. aimed to provide policy evaluation tools using machine learning methods. In their study, they collected 45 privacy policies, and they classified the privacy policy into the specified GDPR categories if they have similar text. These categories are; data collection, protection of children, third-party sharing, data security, data retention, data aggregation, control of data, privacy settings, account deletion, privacy breach notification, policy changes [19]. This tool takes the content from the given privacy policy link and displayed the result for each category to the user. Their system extracts content by using Boilerpipe from the given web page link. They used Naïve Bayes, SVM, Decision Tree, and Random Forest [19] at the prediction step. They calculated average accuracy of the system as 74%.

Sebastian Zimmeck et.al created a website privacy policy corpus as a dataset for a researcher. Their corpus consists of 115 privacy policies and 23K annotations. They defined 10 categories; first-party collection/use, third party sharing, user choice/control, user access, data retention, data security, data security, policy change, do not track, international and specific audiences, other [20]. These categories are used in the annotation tool. This tool helps annotators to create a corpus. They also presented their experience in automating this manual process. Their automatic system tries to predict given privacy policy part. At the prediction, they run SVM, LR, and HMM on this corpus [20]. They showed that the F1 score of the SVM method was slightly higher than the others.

Another study is done in this field is to create automated tools to analyze privacy policies. They stated that in their study all available algorithms on Weka are tested. They implemented Privee as a browser extension to assess the privacy policy completeness of the specified website [21]. This extension takes the content of the privacy policy of the current web page. Then, the system checks the result of this content is available at the repository. If the result is found at the repository, analysis result is shown to the user. On the other hand, if the system cannot find result at this repository, content is analyzed. Firstly, content is sent to rule classifier and machine learning preprocessor. After this process, result is created and displayed on the screen. The success of the system is evaluated with the F1 score. The final F1 score is calculated 90% of this extension.

In another study [29], they proposed a technique which extracts summary from the privacy policies. They summarized the privacy policy considering specified 10 questions related with the privacy and security [29]. They used 400 privacy policies as a corpus. They also presented this technique as a browser extension. This extension also specifies risk level of the content.

In the Usable Privacy Project [30], they extracted relevant key features semi-automatically by using crowdsourcing, machine learning and natural language processing. Then they share these features to the user. These features help the user in decision-making in terms of privacy.

CHAPTER 3

METHODOLOGY

3.1. Corpus Preparation

Web pages links of smart city applications were determined by using keywords "smart city, smart life, etc." via a web crawler, and all found links are collected in a file. The privacy policies of these pages were downloaded and saved to the file simultaneously for consistency on 27 Jan 2020. A total of 68 privacy policies were collected, and only those specific to the web page were eliminated, resulting in 42 policies remaining. Because what we need is that these policies are inclusive of the services and applications. 70% (30) of the collected 42 policies were used to train the system, while the remaining 30% (12) of the documents were used in the testing phase. Related sentences from 42 documents were found by taking into consideration the GDPR elements determined in Figure 2.2.

3.2 Preprocessing

Data preprocessing has an impact to get better results in the machine learning algorithms. It contains many methods to prepare data for the final state. In this step, stop word removal, tokenization and stemming are applied to the training and testing set. At the preprocessing step, stop word removal, tokenization and stemming methods are applied.

The most common words in any language are called stop words. Application goal is vital to determine to remove stop words or not.

Tokenization helps us to split the text into small pieces such as words, subwords, or characters.

Stemming enables finding the root of the word by removing suffixes from it. Porter Stemmer is the most known stemmer method in this area, but it is specific for the English language.

3.3 Probabilistic Relevance Model

Information retrieval systems aim to extract relevant needed information from collections of an information system [22]. There are lots of information retrieval models, in this study probabilistic relevance model is used.

This model is useful to create a ranking function which is used by the ranking method. This model ranks the documents according to their relevance with the given keywords [23]. Okapi BM25 is the well-known ranking function in this category.

3.3.1 Okapi BM25

Okapi Best Matching 25 (BM25), which is generally used by search engines, is a ranking function that lists the most relevant documents according to the given search query. In the automated system, Okapi BM25 [24] is used to find the most relevant sentences from the document. The implementation of BM25 is as:

$$RSV_d = \sum_{t \epsilon q} log\left(\frac{N}{df_t}\right) . \frac{(k+1)tf_{td}}{k\left((1-b)+b\times\left(\frac{|d|}{L_{avg}}\right)\right)+ tf_{td}}$$

Figure 3. 1 BM25 score of a document

In Figure 3.1; N is the total number of documents, d is the document, t is term, df_t is total number of documents included t, tf_{td} is TF value of term t in d, |d| is the document length, L_{avg} is the average length of all documents, k and b are experimental parameters.

The scores after applying BM25 did not produce results as expected. Because of this reason, the new formula is searched and BM25L is found. Lv & Zhai developed this formula because they observed that BM25 fails in large documents. This formula produced a better result in the created system.

$$RSV_d = \sum_{t \epsilon q} log\left(\frac{N+1}{df_{t+0.5}}\right) . \frac{(k+1)c_{td}}{k+ c_{td}}$$

$$\text{where } c_{td} = \frac{tf_{td}}{(1-b)+b\left(\frac{|d|}{L_{avg}}\right)}$$

Final formula is:

$$RSV_d = \sum_{t \epsilon q} log\left(\frac{N+1}{df_{t+0.5}}\right) . \frac{(k+1)(c_{td}+\delta)}{k+ (c_{td}+\delta)}$$

Figure 3. 2 BM25L formula

In Figure 2.2, BM25L formula is the arrangement of BM25 formula. They added positive constant δ to c_{td} to prevent failing the long documents [25]

3.4 TF-IDF

While Term Frequency (TF) measures the number of occurrences of a term in a document, the Inverse Document Frequency (IDF) measures the importance of the word across the documents [26] as stated in Figure 3.3. Especially this method is used in areas of Information Retrieval and Text Mining to identify the most used words in the document.

TF(t) = (Number of occurrences of word w in a document) / (Total number of words in the document) [26].

IDF(t) = log(Total number of documents / Number of documents that includes term t).

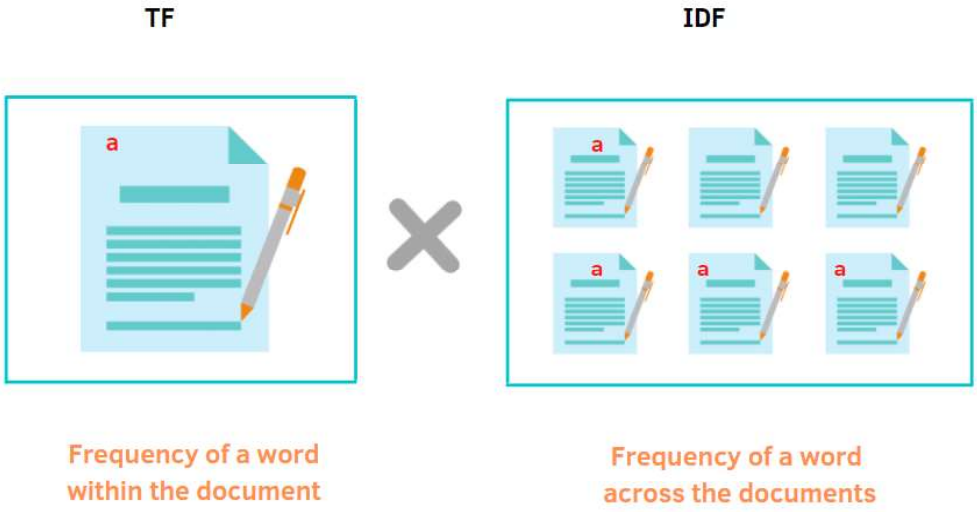

Figure 3. 3 Tf-Idf representation [27]

Then TF-IDF is calculated with the following formula: TF-IDF(t) = TF(t) * IDF(t)

Let's assume we have a document containing 50 words, 5 of them is "paper" word.

TF("paper") => 5 / 50 = 0,1

Furthermore, if we have 1000 documents and "paper" appears 10 of them, IDF("paper") will be log(1000/10) = 4.

Tf-Idf weight of this word = 0,1 * 4 = 0,4

3.6 Classification with Threshold Value

The Okapi BM25 system takes a specified query, then orders sentences, which are related with query, according to score. In this system, most of the documents have a score. Although some of these documents have a score, it is possible that they may be irrelevant. For this reason, by using a threshold value, sentences are cut after a point. This threshold value is found by trial according to your system. The value or values that lead to the most optimal solution are used as the threshold value.

Using the specified threshold value, if the ranked documents are above this threshold value, they are considered to be the most relevant to the given query. The system is executed assuming that the documents below this threshold value are not related to the given query.

CHAPTER 4

IMPLEMENTATION

4.1 Manual Assessment

Privacy policies are legal documents that inform users about the processing of their personal information. Those documents are mandatory in many countries and both websites and applications provide those documents to enable their users to make more informed decisions.

After the EU GDPR became enforceable, we analyzed some smart city applications to observe the effects of it. A privacy policy is a statement or legal document that explains how to use, process, and manage the customer's collected data. For this reason, firstly 68 smart city-related website privacy policies are specified by using the "smart city" keyword and Google engine. These privacy policies are collected on 27 Jan 2020. After the collection of websites, we realized some of the privacy policies are written for Website users. We eliminated these policies, at the final, we only have 42 websites as listed in Table 4.1. According to specified categories, each privacy policy is labeled manually.

Each privacy policy has been analyzed by two independent raters to increase the reliability. In the manual labeling phase, sentences referring to the specified GDPR elements were collected for detailed evaluation.

Table 4. 1 Application List

ABB	Alcatel	AMCS	AppyParking
Aptus	Armis	AT & T	Avectrics
AvePoint	AWS	Axis	Bercman
BigBelly	BlockPower	Cisco	CityMapper
Civic Insight	Compology	FrogParking	Hitachi
Huawei	IBM	Innova	Intel
Itron	Mapillary	Microsoft	Necam
Nokia	Oracle	Panasonic	Philips Lighting
Qognify	Qualcomm	Sentiance	Siemens
Signify	Telcred	TIER	UrbanFootPrint
Volocopter	Coyero		

Explicit Consent

The statements regarding gathering explicit consent are unclear in 11 privacy policies. There are vague statements such as “We use information about you where you have given us consent”. Those policies do not explain how the consent of the data subject was obtained. Three privacy policies express that the consent is gathered via use of their Website or application which does not meet the standard of an unambiguous indication by clear affirmative action in Article 4. Only five privacy policies provided more clear explanations and stated that the consent of the user will be obtained via contract. Two privacy policies emphasize that the consent is obtained through the use of the website, which is confusing given the fact that those policies are written specifically to applications.

Right to Withdraw Consent

Among the 42 privacy policies, 22 policies claim that the data subjects can withdraw consent at any time. However, those policies do not provide any information about how to exercise this right. Two policies acknowledge that when the data subjects decide to withdraw their consent, they should contact them via email.

Right to Erasure

In the majority of the privacy policies (24 policies), explanations regarding this right are limited to vague statements where it is claimed to support this right without any further explanation. Eight privacy policies mention that if the data subject wants their data to be deleted, they should contact them via email. One of the privacy policies express that the data subjects can exercise this right under certain conditions, however, they do not provide further explanation about those conditions. Only one privacy policy acknowledges that the data will be destroyed or anonymized upon request. Perhaps surprisingly, one of the policies states that data subjects can delete some certain information about their accounts which are listed within the privacy policy.

Right to Data Portability

Among the 42 privacy policies, ten policies explain that users can get a copy of their personal data in a structured, commonly used, machine-readable format upon request. Those policies do not provide clear information about how data subjects can request to access their personal information or the format of this information. Only four policies claim that the users have the right to transmit that data to another controller. Two policies cover this right as one of the rights respected by their applications however they do not elaborate it. One policy specifies the data format and states that the users can receive the data in a common format such as csv.

Right of Access

Right of access is claimed to be provided in 33 privacy policies without further explanation. Only one privacy policy claims to provide data subjects with information about whether their personal information is sold or disclosed and with whom and why. One privacy policy covers a surprising statement and expresses that only limited data can be accessed by the users.

Storage Limitation

32 privacy policies do not give specific time related to retention of data. Vague statements such as “certain period of time”, “reasonable period of time”, “minimum period of time based on legislation” or “no longer needed” have been used by those policies. Only one policy gives a specific duration time for personal information.

Transparency

Among 42 privacy policies, 29 policies are transparent about the types of the personal information processed by their systems. One privacy policy gives a list of personal information; however, it also expresses that the personal information collected by their system is not limited to this. 7 privacy policies did not specify what kind of data you collected.

Data Minimization

Our results revealed that there are only 6 policies that cover this principle. Discussions in three policies are limited to vague statements which express that minimum data is collected by the systems. Two policies cover this principle as one of the principles respected by their systems, however, do not provide any further information to explain how they address it. There is one policy which states that users can contact to limit the data their system collects.

Number of policies that mention the investigated GDPR elements are given in Figure 4.1.

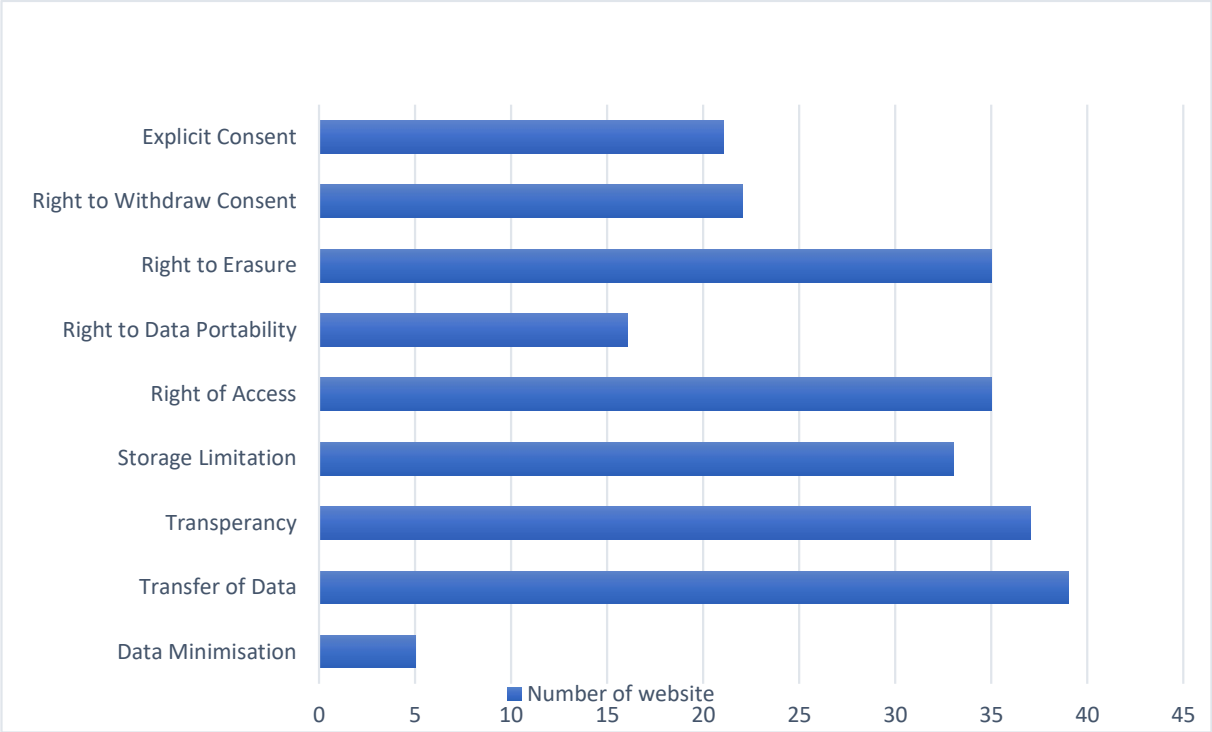

Figure 4. 1 Distribution of the GDPR elements in privacy policies

4.2 Automatization of Privacy Policy Assessment

It is very time-consuming to manually evaluate the compatibility of Smart City privacy policies with GDPR, an automated system was developed to speed up this evaluation process.

4.2.1 Dataset Preparation

Sentences related to each of the 9 elements were manually found by reading the whole documents by taking consider meaning of the elements, and extracted from the collected 42 documents. 30 of them were determined to be used in the training phase and the remaining 12 documents were determined to be used in the testing phase. The sentences extracted from 30 documents were used as input to train the system, while the sentences extracted from 12 documents were used to validate the system. When separating these sets, the following were considered:

- There must be sentences about each element in each set.
- The sentences related to the specified elements should not be too few in the test set.

4.2.2 Preprocessing on the Dataset

At the preprocessing step, stop word removal, tokenization and stemming methods are used for data cleaning.

To remove stop words, python nltk.corpus module is used inside Okapi BM25 and fastText implementations. Removed words are set to the English language. Word tokenization is executed by using nltk.tokenize package at the tokenization step. At the stemming step, SnowballStemmer is used inside nltk.stem.snowball package.

4.2.3 How to Label Data

Data labeling plays an important role before training the system. With this methodology, data will have context for clear identification. Data labeling helps to train the system more accurately. Labeled data is crucial for supervised learning. Before the testing data, the train set is labeled with specified elements.

How to Label Data for Document-Based Implementation

The sentences extracted for each GDPR element from 30 documents were collected in the file of the relevant element. All sentences related to the right to erasure were collected into a right to erasure file and this step was repeated for every 9 elements. As a result, we have 9 separate documents. At the final, total 238 text passages are extracted for the training set. Their distribution is illustrated in the Figure 4.2. These documents are ready for the training.

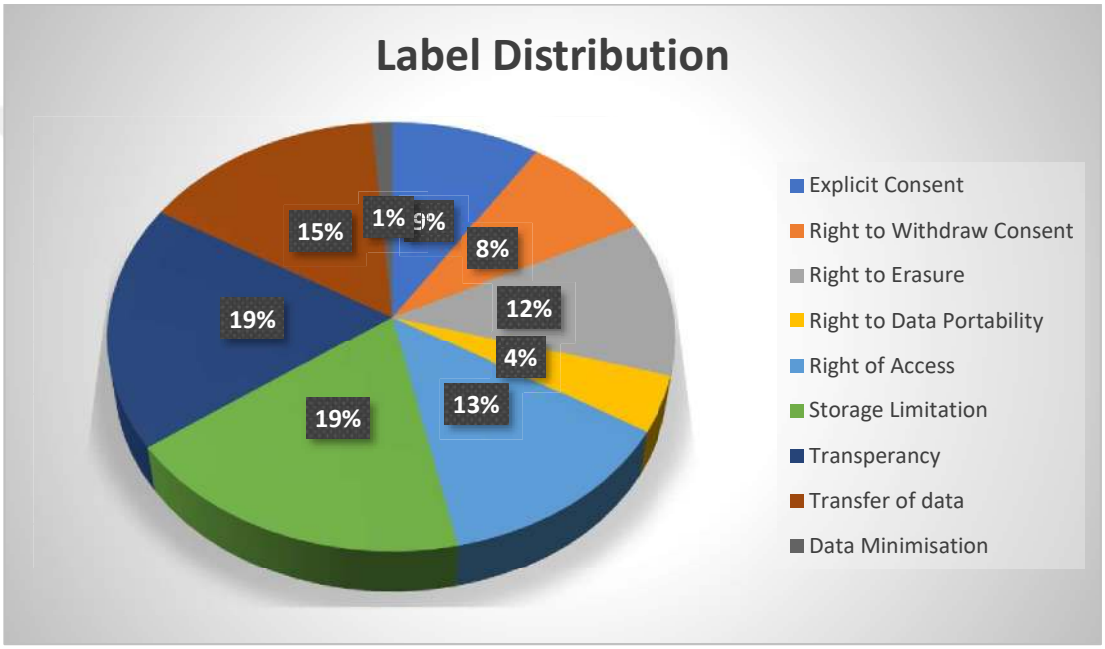

Figure 4. 2 Label distribution of document-based implementation

How to Label Data for Sentence-Based Implementation

The extracted 238 passages at the document-based labeling phase, also they are used in this implementation. The related sentences for each element, are created at the corpus preparation phase. At the sentence-based implementation, it is needed to add one more label to classify sentences which do not belong to any of the categories. This label is specified as an “other”. At the final, total 299 text passages are extracted for the training set. Their distribution is illustrated in the Figure 4.3.

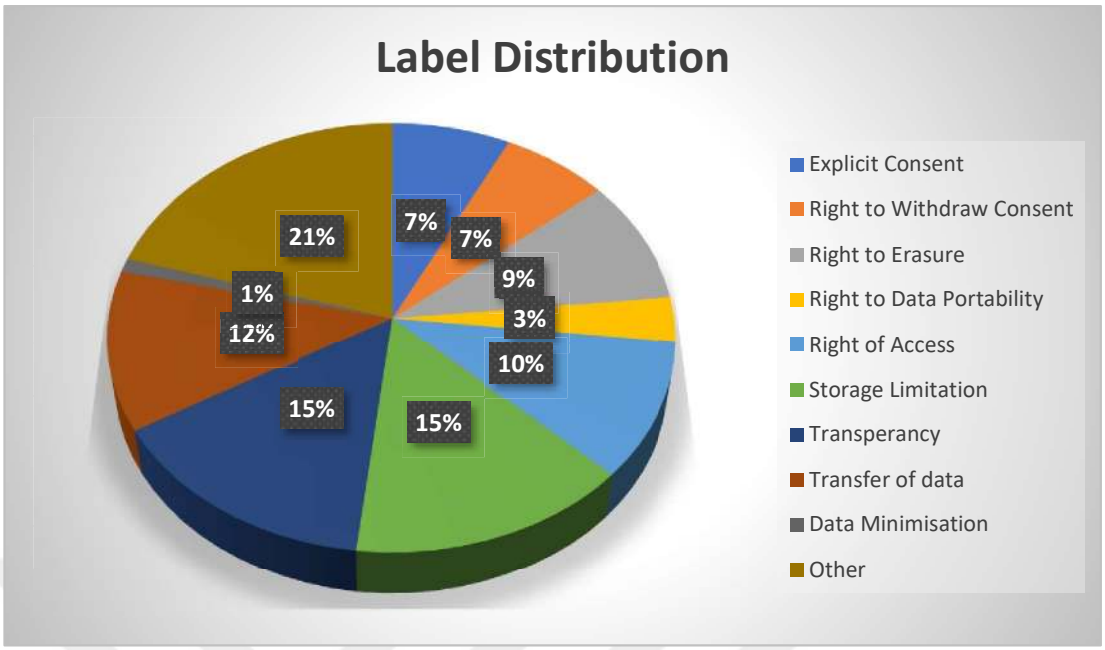

Figure 4. 3 Label distributions of sentence-based implementation

4.2.4 Implementation with Okapi BM25

4.2.4.1 How to Determine Query Words

The query words of each GDPR element were found by the term frequency – inverse document frequency (TF-IDF) method over labeled 9 documents which are extracted from 30 documents.

At the Tf-Idf phase, sklearn.feature_extraction.text submodule is used inside python library. This library includes TfidfVectorizer class inside it.

Document-term matrix is created by execution of fit_transfrom method in this class. After execution of this method, our keywords are ranked according to their point for each element. This list was shortened from the point where the irrelevant word order started.

Right to Erasure Keywords

For the right to erasure document, extracted keywords are listed in Table 4.2.

Table 4. 2 Right to erasure keywords

	Term	Score
0	delet	33,86294
1	correct	17,24662
2	access	15,10826
3	updat	13,41404
4	eras	8,815891
5	remove	6,772589

Right of Access Keywords

For the right of access document, extracted keywords are listed in Table 4.3.

Table 4. 3 Right of access keywords

	Term	**Score**
0	access	25,68404
1	correct	19,16291
2	hold	15,42781
3	copi	15,33033
4	delet	11,85203
5	updat	11,49774
6	held	7,828314
7	com	7,828314
8	suppress	6,611918
9	review	6,611918

Data Minimisation Keywords

For the data minimisation document, extracted keywords are listed in Table 4.4.

Table 4. 4 Data minimization keywords

	Term	Score
0	suppli	5,218876
1	limit	4,407946
2	exampl	3,386294
3	servic	3,021651
4	sure	2,609438
6	absolut	2,609438
7	inaccur	2,609438
8	stop	2,203973
9	relev	1,916291

Storage Limitation Keywords

For the storage limitation document, extracted keywords are listed in Table 4.5.

Table 4. 5 Storage limitation keywords

	Term	Score
0	retain	57,30329
1	long	36,53213
2	necessari	34,49323
3	period	28,65165
4	retent	23,48494
5	longer	22,03973
6	compli	19,83576
7	busi	19,64073
8	disput	18,26607
9	collect	16,61908
10	account	16,61908
11	onli	13,54518
12	resolv	13,04719
13	year	13,04719

Right to Data Portability Keywords

For the right to portability document, extracted keywords are listed in Table 4.6.

Table 4. 6 Right to portability keywords

	Term	Score
0	format	8,815891
1	copi	7,665163
2	machin	6,611918
3	readabl	6,611918
4	receiv	5,748872
5	common	4,407946
6	structur	2,609438
7	port	2,609438
8	portabl	2,609438
9	anoth	2,609438
10	csv	2,609438
11	machine	2,609438

Right to Withdraw Consent Keywords

For the right to withdraw consent document, extracted keywords are listed in Table 4.7.

Table 4. 7 Right to withdraw consent keywords

	Term	Score
0	consent	37,24924
1	withdraw	33,05959
2	time	12,0866
3	basi	7,828314
4	affect	7,828314
5	chang	3,832581
6	base	3,386294
7	maintain	3,021651
8	declar	2,609438
9	alreadi	2,609438
10	revok	2,609438

Transfer of Data Keywords

For the transfer of data document, extracted keywords are listed in Table 4.8.

Table 4. 8 Transfer of data keywords

	Term	Score
0	transfer	44,02183
1	parti	42,32868
2	share	28,74436
3	protect	27,19486
4	servic	24,17321
5	european	15,65663
6	locat	15,42781
7	jurisdict	15,33033
8	busi	15,10826
9	compli	13,22384
10	outsid	13,22384

Transparency Keywords

For the transparency document, extracted keywords are listed in Table 4.9.

Table 4. 9 Transparency keywords

	Term	Score
0	address	68,32316
1	collect	66,47633
2	devic	52,18876
3	servic	45,32477
4	includ	39,28147
5	number	38,94239
6	site	38,32581
7	locat	26,44767
8	type	26,09438
9	product	25,39721

Explicit Consent Keywords

For the explicit consent document, extracted keywords are listed in Table 4.10.

Table 4. 10 Explicit consent keywords

	Term	Score
0	consent	27,09035
1	websit	9,581454
2	contract	8,815891
3	servic	7,554128
4	unit	6,611918
5	contractu	6,611918
6	collect	6,043302

4.2.4.2 Prediction with Test Set

The words belonging to each element obtained during the training phase are used as search keywords at this stage, and Okapi BM25 is run on the test documents.

How to determine threshold value

For a specified document, the average scores of the second-order sentences from each query is accepted as the first threshold. In order to avoid sentences that could escape this threshold value, the average score of the sentences in the fifth-order is also accepted as the second threshold value. Let's assume we executed Okapi BM25

with right to erasure keywords for a first test document. After the execution, each sentence has greater score than the first threshold is specified. If there is no sentence with this threshold value, this document is evaluated by considering the 2nd threshold value at this time.

As a summary;

$$1st\ threshold = \frac{the\ sum\ of\ the\ scores\ of\ the\ sentences\ in\ the\ second\ order\ from\ each\ query}{9}$$

$$2nd\ threshold = \frac{the\ sum\ of\ the\ scores\ of\ the\ sentences\ in\ the\ fifth\ order\ from\ each\ query}{9}$$

Since the sentences below the threshold for each element can be labeled as other, the "other" label is not used separately.

How to test the system

12 documents specified for the test case are listed in Table 4.11

Table 4. 11 Test case document list

Avectric	CityMapper	Necam
AWS	Huawei	Signify
Bercman	Mapillary	Tier
BigBelly	Microsoft	Volocopter

For each GDPR element related query words, sentences are tried to extract with Okapi BM25 implementation. This system takes the given document which passed through preprocessing, then splits to the sentences by using regular expressions. If the system finds any sentences above the threshold value, it accepts that this sentence related with specified GDPR element. If score of the sentences is below the threshold values, these are accepted that are not related.

4.2.5 Implementation with fastText

Facebook Research Team developed a new model for classification called as a fastText which is a word embedding model [31]. This model is created by taking consider n-gram, hierarchical softmax classifier, dimensionality reduction [32]. They shared their experiments with fastText library. It is an open-source library which is used to classify texts and word representations. This method is used to compare it with the Okapi-BM25 method. Python fastText module is for the Mac and Linux systems. Because of this reason, wrapper .whl file is used to install it on Windows machines. It has a dependency with Numpy& Scipy and pybind11 modules. At the training step, the train_supervised method is used inside it. This method takes a labeled text file as an input. This text file includes sentences inside it as the Figure 4.4.

```
__label__dataMinimisation Sentence1

__label__dataMinimisation Sentence1

__label__dataAccessRight Sentence2

__label__dataPortability Sentence3

__label__other Sentence4
```

Figure 4. 4 fastText training input format

Specification file of fastText proposed few steps to get better modeling. These following steps are applied to train the model:

- Preprocessing on data,
- Epoch rate is used,
- Learning rate is used,
- Finally, word n-grams value is applied.

All implementations are written with python by using fastText module. The model takes epoch rate, learning rate, and word n-gram parameters for better prediction at this step. The parameters of the implementation can be listed as follows:

- Epoch rate is assumed as a 25,
- Learning rate is assumed as a 1.0
- Word n-gram is assumed as a 2

These parameters have been decided by taking consider the specification of the fastText. As it is a classification algorithm, a threshold value was not used. After training the system, the system is ready to predict the given documents. Firstly, data preprocessing is executed for the selected documents belong to test case. Then, these documents are splitted to sentences by using regular expressions, and these sentences are predicted with given 10 labels. At the evaluation phase, only 9 label results are used, the “other” label is not taken considered. Because the system only needs sentences related with 9 elements.

4.3 Evaluation metric

In the document-based evaluation, precision, recall, and accuracy values are found and the accuracy value is used as a success criterion. Precision, recall and F1 values were found in the sentence-based evaluation stage, and the F1 score is used as the evaluation criterion, since the unbalanced distribution of the classified sentences at this stage was taken into account.

4.4 Implementation Results

4.4.1 Document Based Result

For each document, implementation result is given in the tables in Appendix A with specified columns in Figure 4.5. In this table has the following information inside it:

- # of Sentences: It indicates sentences count which has higher score than the threshold. If the sentences count is zero, the system shows that document has not included related element. If the sentences count is greater than zero, system shows that document has related element.
- Actual means, the document has or has not related elements inside it in reality.
- For the assessment column the following rules are executed:
  1. If the document includes specified element and system finds sentences, in addition, anyone of founded sentences is related with this element, prediction is marked as a positive.
  2. If document does not include specified element and system cannot find any sentence, prediction is marked as negative.
  3. If document does not include specified elements and system finds sentences, prediction is marked as a positive.
  4. If document includes specified element and system finds sentences, but none of them is related with this element, system is marked as a negative.
- First Threshold: It indicates whether first threshold is used or not.
- Second Threshold: It indicates whether second threshold is used or not.

# of Sentences	Actual	Prediction	First Threshold	Second Threshold

Figure 4. 5 Document assessment columns

After the evaluation of all documents, table is created for every element to show TP, TN, FP and FN values. Results are stated in Table 4.12.

Table 4. 12 Document-based Okapi BM25 result

	TP	TN	FP	FN
Right to erasure	9	0	0	3
Right to withdraw consent	10	0	2	0
Transfer of data	12	0	0	0
Data Portability	6	5	0	1
Explicit Consent	7	0	1	4
Storage Limitation	11	0	1	0
Data Minimisation	1	0	10	1
Right of access	6	0	0	6
Transparency	12	0	0	0

Precision, recall, and F1 score are calculated by using the confusion matrix for each element in Table 4.13.

Table 4. 13 Precision, recall, and F1 score of Okapi BM25 implementation

	Precision (%)	Recall (%)	F1
Right to erasure	1.0	75	85,71
Right to withdraw consent	83,3	100	90,88
Transfer of data	1.0	1.0	100
Data Portability	100	85,7	92,2
Explicit Consent	87,5	63,6	73,65
Storage Limitation	91,6	100	95,61
Data Minimisation	9,09	50	15,17
Right of access	1.0	50	66,6
Transparency	1.0	1.0	100

When we evaluate these results, especially right of access and data minimization elements have lower score than others. For the data minimization element, documents do not include enough data minimization sentences inside it. Because of this reason, it is not impossible to get higher score. For the right of access, the problem is that keywords are extracted from the training set are very common world. This situation causes to extract wrong sentences from the document.

If we calculate accuracy for the overall system, we have total TP is 74 cases, total FN is 15 cases, total TN is 5 cases, total FP is 14 cases. Results are stated as follows:

- accuracy is calculated as 73.17%.
- Precision of the system is calculated as 83.1%,
- Recall of the system is calculated as 82.2%,
- F1 Score is calculated as 82.64 %.

These results show that this system has an ability to extract GDPR elements from the document with 82,64% F1 score. These calculations show that the accuracy of the system is not enough to specify system performance. F1 score gives a better result than accuracy.

4.4.2 Sentence-Based Results

In this assessment, calculations are done with sentence-based classification. All details are given the following titles. For the sentence-based result, total 2844 sentences are tried to classify correctly by the implementations. The distributions of these sentences are illustrated in the Figure 4.6.

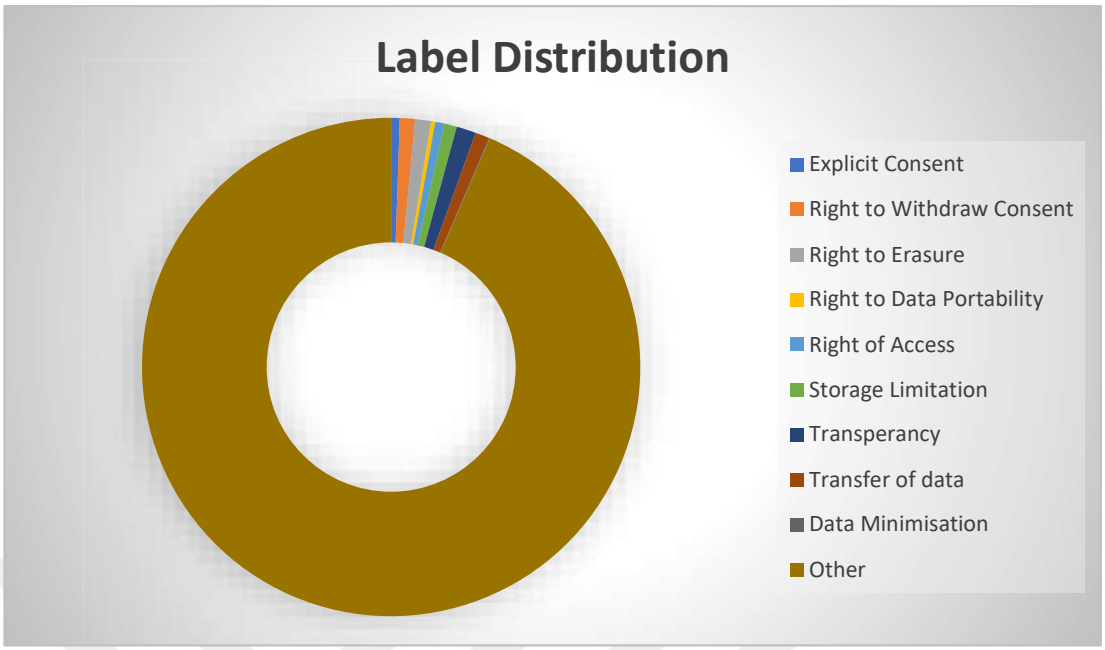

Figure 4. 6 Distributions of labeled test case sentences

4.4.2.1 Sentence-Based Results with Okapi BM25

Sentences which are above the threshold values are assumed they are related with specified query. Sentences which are below the threshold values are assumed they are not related with this query. All calculations are briefly stated in table 4.14.

After filling the TP, TN, FP and FN values, precision, recall, F1 score, and accuracy are calculated by using these values. The precision of the system is calculated as 55.1%, recall of the system is 72.4%, the F1 score is calculated as 62.6%, and accuracy is calculated as 99.3%.

Table 4. 14 Sentenced-based Okapi BM25 result

	TP	TN	FP	FN
Right to erasure	18	2804	11	11
Right to withdraw consent	12	2805	11	16
Transfer of data	23	2805	13	3
Data Portability	7	2834	2	1
Explicit Consent	11	2810	18	5
Storage Limitation	16	2810	11	7
Data Minimisation	1	2823	19	1
Right of access	12	2814	13	5
Transparency	34	2797	11	2

4.4.2.2 Sentence-Based Results with fastText

Although there were 10 labels at this stage, a confusion matrix was created out of 9 of them. Because the 10th label was created only to distinguish sentences that do not belong to a certain category. Results are stated in Table 4.15.

Table 4. 15 Sentence-based fastText result

	TP	TN	FP	FN
Right to erasure	24	2801	14	5
Right to withdraw consent	20	2796	20	8
Transfer of data	25	2791	27	1
Data Portability	6	2829	7	2
Explicit Consent	14	2805	23	2
Storage Limitation	22	2771	30	1
Data Minimisation	1	2833	9	1
Right of access	15	2798	29	2
Transparency	34	2775	33	2

Using this confusion matrix belongs to fastText, precision, recall, F1 score, and accuracy are calculated. The precision of the system is calculated as 45.6%, recall of the system is 87.0%, the F1 score is calculated as 59.8%, and accuracy is calculated as 99.1%.

4.5 Comparison of Okapi BM25 with fastText

When fastText and Okapi BM25 implementations are compared, especially F1 scores are taken into account. With a slight difference, the F1 score of the Okapi BM25 implementation is higher than the fastText implementation. When we compare the results of these implementations, fastText implementation has more FP values than Okapi BM25 implementation. On the other hand, precision score of the Okapi BM25 is higher than fastText implementation.

CHAPTER 5

CONCLUSION, LIMITATIONS AND FUTURE WORK

This study covers the results of a data-driven study conducted on privacy policies of smart city applications in order to address how the challenges introduced by GDPR have been addressed by those systems. We have focused on smart city applications considering the huge amount of personal data they process with the help of recent developments in IoT and Big Data. Our results reveal that most systems did not provide explicit acknowledgement about GDPR. Several GDPR elements such as data minimization or consent have been overlooked by those policies. It is possible to conclude that the status of the privacy policies is not satisfactory considering the sensitivity and the variety of the personal information processed by those systems. This is clearly an area in need of more research.

In this study also provides a system to assess documents automatically by taking consider the specified 9 elements. When we compare the Okapi BM25 system with fastText, Okapi BM25 shows better performance than fastText. On the other hand, this system performance can be increased with modification TF-IDF step for Okapi BM25 implementation. In this step, if we can extract more proper words, it will affect the F1 score of the system. Also, this system is now suitable for English documents. Besides this, at the evaluation of the training case, if the system can extract the meaning of the sentences, the system can accelerate the assessment.

In the future, this system can serve for most of the language documents. Also, data can be increased to train the system. Modification Okapi BM25 step by taking into consideration the meaning of the sentences can help to shorten the assessment step. Furthermore, it is also planned to compare the Okapi BM25 method with other deep learning and machine learning methods.

REFERENCES

[1] Arasteh, H., Hosseinnezhad, V., Loia, V., Tommasetti, A., Troisi, O., Shafie-khah, M., & Siano, P., "Iot-based smart cities: A survey", 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), 2016.

[2] Weber M., Lučić D. & Lovrek I., "Internet of Things context of the smart city", 2017 International Conference on Smart Systems and Technologies (SST), 2017.

[3] Woetzel, J., Remes, J., Boland, B., Lv, K., Sinha, S., Strube, G., et al., "SMART CITIES: DIGITAL SOLUTIONS FOR A MORE LIVABLE FUTURE", Mc Kinsey Global Institute, 2018.

[4] Wray, S., "SmartCitiesWorld", Retrieved March 20, 2020, from https://www.smartcitiesworld.net/news/news/2020-expected-to-see-almost-20-increase-in-smart-city-spending--5028.

[5] Van Zoonen, L., "Privacy concerns in smart cities", Government Information Quarterly, pp. 472-480, 2016.

[6] Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I., & Toft, T, "Privacy-preserving face recognition", In International symposium on privacy enhancing technologies symposium, 2009.

[7] Hasan, A. S. M. T., Jiang, Q. & Li, C., "An Effective Grouping Method for Privacy-Preserving Bike Sharing Data Publishing", Future Internet, 9 (4), p. 65, 2017.

[8] Folianto F., Low, Y. S. & Yeow, W. L., "Smartbin: Smart waste management system," In 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015.

[9] Al Nuaimi, E., Al Neyadi, H., Mohamed, N. & J. Al-Jaroodi, "Applications of big data to smart cities," Journal of Internet Services and Applications, 6 (1), p. 25, 2015.

[10] Martinez-Balleste, A., Solanas, A. & Perez-martinez, P. A., "The pursuit of citizens' privacy: a privacy-aware smart city is possible," IEEE Communications Magazine, 51(6), pp. 136-141, 2013.

[11] Zhang, K., Ni, J., Yang, K., Liang, X., Ren, J. & Shen, X., "Security and Privacy in Smart City Applications: Challenges and Solutions", IEEE Communications Magazine, pp. 122-129, 2017.